Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/JP05/005440

International filing date: 17 March 2005 (17.03.2005)

Document type: Certified copy of priority document

Document details: Country/Office: JP
Number: 2004-079077
Filing date: 18 March 2004 (18.03.2004)

Date of receipt at the International Bureau: 07 April 2005 (07.04.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)

World Intellectual Property Organization (WIPO) - Geneva, Switzerland
Organisation Mondiale de la Propriété Intellectuelle (OMPI) - Genève, Suisse

PCT/JP2005/005440

17.3.2005

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００４年　３月１８日

出願番号
Application Number: 特願２００４－０７９０７７

[ST. 10/C]: ［ＪＰ２００４－０７９０７７］

出願人
Applicant(s): 日本電気株式会社

２００４年　８月３０日

特許庁長官
Commissioner,
Japan Patent Office

小川洋

出証番号　出証特２００４－３０７７５８４

【書類名】　特許願
【整理番号】　34403346
【提出日】　平成16年 3月18日
【あて先】　特許庁長官　殿
【国際特許分類】　G06F 17/30
G06F 9/44
G06F 17/27
【発明者】
【住所又は居所】　東京都港区芝五丁目７番１号　日本電気株式会社内
【氏名】　坂尾　要祐
【発明者】
【住所又は居所】　東京都港区芝五丁目７番１号　日本電気株式会社内
【氏名】　佐藤　研治
【発明者】
【住所又は居所】　東京都港区芝五丁目７番１号　日本電気株式会社内
【氏名】　赤峯　享
【特許出願人】
【識別番号】　000004237
【氏名又は名称】　日本電気株式会社
【代理人】
【識別番号】　100080816
【弁理士】
【氏名又は名称】　加藤　朝道
【電話番号】　045-476-1131
【手数料の表示】
【予納台帳番号】　030362
【納付金額】　21,000円
【提出物件の目録】
【物件名】　特許請求の範囲　1
【物件名】　明細書　1
【物件名】　図面　1
【物件名】　要約書　1
【包括委任状番号】　9304371

【書類名】特許請求の範囲
【請求項１】
入力した文書から文構造を作成する手段と、
前記文構造の部分構造に対して、予め定められた所定の変換操作を行うことで、前記部分構造と意味の類似したパターンの類似構造を作成する手段と、
前記意味の類似したパターンを同一パターンと判定してパターン検出を行う手段と、
を備えている、ことを特徴とするテキストマイニング装置。
【請求項２】
テキストマイニングの対象となる文書の集まりを記憶する記憶部と、
前記記憶部の前記文書を入力して解析し文構造を取得する解析部と、
を備え、
前記解析部は、前記文書を解析し、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした文構造を生成する、ことを特徴とする請求項１に記載のテキストマイニング装置。
【請求項３】
前記類似構造を生成する手段が、
前記文構造について並列変形を行う手段と、
前記文構造の部分構成を生成する手段と、
前記文構造及び／又は部分構造の有向枝の無向枝化を行う手段と、
同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う手段と、
前記文構造及び／又は部分構造における順序木の無順序木化を行う手段と、
を備え、
前記類似構造を前記文構造の部分構造の同値類とする、ことを特徴とする、請求項１に記載のテキストマイニング装置。
【請求項４】
テキストマイニングの対象となる文書の集まりを記憶する記憶部と、
前記記憶部から前記文書を読み出して解析し文構造を取得する解析部と、
前記解析部により解析して得られる文構造の部分構造に対して予め定められた所定の変形操作を行い、意味的に類似したパターンの類似構造を生成する類似構造生成部と、
前記類似構造生成部によって生成された類似構造を、生成元の部分構造の同値類として扱いパターン検出を行うパターン検出部と、
を備えている、ことを特徴とするテキストマイニング装置。
【請求項５】
前記パターン検出部は、前記類似構造を、生成元の部分構造の同値類として扱い頻出パターンを検出する、ことを特徴とする請求項４に記載のテキストマイニング装置。
【請求項６】
前記類似構造生成部が、
前記文構造について並列変形を行う手段と、
前記文構造の部分構成を生成する手段と、
前記文構造及び／又は部分構造の有向枝の無向枝化を行う手段と、
同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う手段と、
前記文構造及び／又は部分構造における順序木の無順序木化を行う手段と、
を備え、
前記文構造の類似構造を生成し、前記類似構造を同値類とする、ことを特徴とする、請求項４に記載のテキストマイニング装置。
【請求項７】
使用者がどこまで類似したパターンを同一と判定してパターン検出を行うか調整する手段を備えている、ことを特徴とする請求項４に記載のテキストマイニング装置。
【請求項８】
テキストマイニングの対象となる文書の集まりを記憶する記憶部と、

前記記憶部から前記文書を読み出して解析し文構造を取得する解析部と、
使用者の入力から文構造の差異の種別ごとに同一構造と判定するか否かを指定する第１の指定項目を生成する類似構造生成調整部と、
使用者の入力から属性値の差異の種別ごとに同一構造と判定するか否かを指定する第２の指定項目を生成する類似構造判定調整部と、
前記類似構造生成調整部によって生成された第１の指定項目に従い、前記解析部で得られた文構造の部分構造に対して所定の変換操作を行い、前記部分構造と意味的に類似した類似構造を生成する類似構造生成部と、
前記類似構造生成部によって生成された類似構造を生成元の部分構造の同値類として扱い、前記類似構造判定調整部の第２の指定項目に従い、属性値の差異を無視しながら、頻出パターンの検出を行う類似パターン検出部と、
を備えている、ことを特徴とするテキストマイニング装置。
【請求項９】
前記解析部は、前記文書を解析し、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした前記文構造を生成し、
前記属性値は、前記文構造に付加された表層格及び／又は付属語情報を含む、ことを特徴とする請求項８に記載のテキストマイニング装置。
【請求項１０】
前記類似パターン検出部は、頻出の類似パターンを検出する、ことを特徴とする請求項８に記載のテキストマイニング装置。
【請求項１１】
前記類似構造生成部が、
前記第１の指定項目に、並列変形の指定がある場合、前記文構造について並列変形を行う手段と、
前記文構造の部分構造を生成する手段と、
前記第１の指定項目に、有向枝の無向枝化の指定がある場合に、前記文構造及び／又は部分構造の有向枝の無向枝化を行う手段と、
前記第１の指定項目に、同義語の置換の指定がある場合、同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う手段と、
前記第１の指定項目に、順序木の無順序木化の指定がある場合、前記文構造及び／又は部分構造における順序木の無順序木化を行う手段と、
を備え、
前記文構造の類似構造を生成し、前記類似構造を同値類とする、ことを特徴とする、請求項８に記載のテキストマイニング装置。
【請求項１２】
入力した文書から文構造を作成する工程と、
前記文構造の部分構造に対する所定の変換操作を行うことで、前記部分構造と意味の類似したパターンの類似構造を作成する工程と、
前記意味の類似したパターンを同一パターンと判定してパターン検出を行う工程と、
を含む、ことを特徴とするテキストマイニング方法。
【請求項１３】
テキストマイニングの対象となる文書の集まりを記憶する記憶部から前記文書を入力して解析し、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした文構造を生成する工程を含む、ことを特徴とする請求項１２に記載のテキストマイニング方法。
【請求項１４】
前記類似構造を生成する工程が、
前記文構造について並列変形を行う工程と、
前記文構造の部分構造を生成する工程と、
前記文構造及び／又は部分構造の有向枝の無向枝化を行う工程と、

同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う工程と、
前記文構造及び／又は部分構造における順序木の無順序木化を行う工程と、
を含み、
前記類似構造を前記部分構造の同値類とする、ことを特徴とする請求項１２に記載のテキストマイニング方法。
【請求項１５】
テキストマイニングの対象となる文書の集まりを記憶する記憶部より前記文書を解析して文構造を取得する工程と、
前記文構造の部分構造に対して予め定められた所定の変形操作を行い、意味的に類似したパターンを有する類似構造を生成する工程と、
前記生成された類似構造を、生成元の部分構造の同値類として扱いパターン検出を行う工程と、
を含む、ことを特徴とするテキストマイニング方法。
【請求項１６】
前記類似構造を、生成元の部分構造の同値類として扱い頻出パターンを検出する工程を含む、ことを特徴とする請求項１５に記載のテキストマイニング方法。
【請求項１７】
前記類似構造を生成する工程が、
前記文構造について並列変形を行う工程と、
前記文構造の部分構造を生成する工程と、
前記文構造及び／又は部分構造の有向枝の無向枝化を行う工程と、
同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う工程と、
前記文構造及び／又は部分構造における順序木の無順序木化を行う工程と、
を含み、
前記文構造の類似構造を生成し、前記類似構造を同値類とする、ことを特徴とする、請求項１５に記載のテキストマイニング方法。
【請求項１８】
使用者がどこまで類似したパターンを同一と判定してパターン検出を行うか調整する工程を備えている、ことを特徴とする請求項１７に記載のテキストマイニング方法。
【請求項１９】
テキストマイニングの対象となる文書の集まりを記憶する記憶部より前記文書を解析して文構造を取得する工程と、
使用者の入力から文構造の差異の種別ごとに同一構造と判定するか否かを指定する第１の指定項目を生成する工程と、
使用者の入力から属性値の差異の種別ごとに同一構造と判定するか否かを指定する第２の指定項目を生成する工程と、
前記生成された第１の指定項目に従い、前記解析部で得られた文構造の部分構造に対して所定の変形操作を行い、前記部分構造と意味的に類似した類似構造を生成する工程と、
前記生成された類似構造を生成元の部分構造の同値類として扱い、前記第２の指定項目に従い、属性値の差異を無視してパターンの検出を行う工程と、
を含む、ことを特徴とするテキストマイニング方法。
【請求項２０】
前記文構造を取得する工程が、文節が節点をなし、少なくとも係り受け関係を係り元の節点から係り先への節点の有向枝で表わした前記文構造を生成し、
前記属性値は、前記文構造に付加された表層格及び／又は付属語情報を含む、ことを特徴とする請求項１９に記載のテキストマイニング方法。
【請求項２１】
前記頻出の類似パターンを検出する、ことを特徴とする請求項１９に記載のテキストマイニング方法。
【請求項２２】

前記類似構造を生成する工程が、
前記第１の指定項目に、並列変形の指定がある場合、前記文構造について並列変形を行う工程と、
前記文構造の部分構造を生成する工程と、
前記第１の指定項目に、有向枝の無向枝化の指定がある場合に、前記文構造及び／又は部分構造の有向枝の無向枝化を行う工程と、
前記第１の指定項目に、同義語の置換の指定がある場合、同義語辞書を参照して前記文構造及び／又は部分構造中の同義語の置換を行う工程と、
前記第１の指定項目に、順序木の無順序木化の指定がある場合、前記文構造及び／又は部分構造における順序木の無順序木化を行う工程と、
を含み、
前記文構造の類似構造を生成し、前記類似構造を同値類とする、ことを特徴とする、請求項１９に記載のテキストマイニング方法。

【請求項２３】
テキストマイニング装置を構成するコンピュータに、
テキストマイニングの対象となる文書の集まりを記憶する記憶部の前記文書を解析して文構造を取得する処理と、
前記文構造の部分構造に対して所定の変換操作を行い、前記部分構造と意味的に類似した類似構造を生成する処理と、
前記生成された類似構造を、生成元の部分構造の同値類として扱い、所定のパターン検出を行う処理と、
を実行させるプログラム。

【請求項２４】
テキストマイニング装置を構成するコンピュータに、
テキストマイニングの対象となる文書の集まりを記憶する記憶部の前記文書を解析して文構造を取得する処理と、
前記文構造の部分構造に対して予め定められた所定の変換操作を行い、前記部分構造と意味的に類似したパターンの類似構造を生成する処理と、
前記生成された類似構造を生成元の部分構造の同値類として扱い、属性値の差異を無視しながらパターン検出を行う処理と、
を実行させるプログラム。

【請求項２５】
テキストマイニング装置を構成するコンピュータに、
テキストマイニングの対象となる文書の集まりを記憶する記憶部の前記文書を解析して文構造を取得する処理と、
使用者の入力から、前記文構造の差異の種別ごとに、同一構造と判定するか否かを指定する第１の指定項目と、属性値の差異の種別ごとに同一構造と判定するか否かを指定する第２の指定項目を生成する処理と、
前記文構造の差異の種別ごとに同一構造と判定するか否かを指定する前記第１の指定項目に従って、前記文構造の部分構造に対して所定の変換操作を行い、意味的に類似したパターンの類似構造を生成する処理と、
生成された類似構造を生成元の部分構造の同値類として扱い、属性値の差異の種別ごとに同一構造と判定するか否かを指定する前記第２の指定項目に従って、属性値の差異を無視しながら頻出パターンの検出を行う処理と
を実行させるプログラム。

【書類名】明細書
【発明の名称】テキストマイニング装置、その方法及びプログラム
【技術分野】
【０００１】
本発明は、構文解析などを用いて、コンピュータ上に蓄積される電子化テキストを構造化して分析を行うテキストマイニング装置、テキストマイニング方法及びテキストマイニング用プログラムに関し、特に、意味の類似した文の構造を同一の構造と判定して分析を行うことができるテキストマイニング装置、テキストマイニング方法及びテキストマイニング用プログラムに関する。
【背景技術】
【０００２】
テキストマイニング装置の一例として、図１０に示すような構成が知られている（後記特許文献１参照）。図１０に示すように、この従来のテキストマイニング装置は、基本辞書記憶部と、文書データ記憶部と、分野依存辞書記憶部と、言語特徴分析装置と、言語解析装置と、パターン抽出装置と、頻出パターン表示装置とを備えている。
【０００３】
図１０に示した従来のテキストマイニング装置は、概略、つぎのように動作する。まず、言語特徴分析装置によって基本辞書と文書データとから分野依存辞書を作成し、言語解析装置によって基本辞書と分野依存辞書と文書データから構文木等の構造を作成し、パターン抽出装置が、この構造を用いて頻出パターンを抽出し、この頻出パターンに合致する文書データ中の文書を、頻出パターン適合文書記憶部に記憶させると同時に、この頻出パターンを出力する。
【０００４】
一般的に、言語解析装置によって作成される構造として、例えば、
（Ａ１）文中の文節を、構造の節点で表し、
（Ａ２）付属語情報を、節点の属性値で表し、
（Ａ３）係り受け関係を、係り元の節点から係り先の節点への有向枝で表し、
（Ａ４）表層格の情報を、有向枝の属性値で表す
という構造が良く用いられる。
【０００５】
ここで、付属語情報とは、付属語によって文節に付加される、進行や完了などの時制、容易や困難などのモダリティ、及び否定などの付属的な概念である。図１１に、この形式で表された「彼は車種Ａが価格を下げたのを知らない」という文の構文構造の一例を示す。文の文節、「彼」、「車種Ａ」、「価格」、「下げる」、「知る」を節点で表わし、付属語情報を、節点の属性値で表し（節点「知る」の属性値として、付属語情報：否定）、係り受け関係を、係り元の節点から係り先への有向枝で表わし（例えば「彼」→「知る」）、表層格の情報を、有向枝の属性値で表している（例えば「彼」→「知る」の有向枝の属性値として「表層格：は」）。
【０００６】
また、構造中のこれらの情報は、全て属性値を持たないラベル付きの節点と、属性値を持たない有向枝のみからなる構造で表現することも可能である。図１２に、この形式で表された「彼は車種Ａが価格を下げたのを知らない」という文の構文構造の例を示す。文の文節「彼」、「車種Ａ」、「価格」、「下げる」、「知る」を属性値を持たないラベル付きの節点で表わし（例えば節点「彼」には「表層格：は」のラベルが付加され、「下げる」には、ラベル「付属語情報：完了」、「表層格：を」が付加されている）、係り元の節点から係り先への有向枝は属性値を持たない有向枝とされる。
【０００７】
【特許文献１】特開２００１－８４２５０号公報（第４、５頁、第３図）
【発明の開示】
【発明が解決しようとする課題】

【０００８】
上記した従来のシステムは下記記載の問題点を有している。なお、以下の問題点及びその解析は、本願発明者らによる研究・検討結果に基づくものであり、図１３、図１４の内容は、問題の在り処を具体的に説明するために、本願発明者らが提示したものである。
【０００９】
第１の問題点は、頻出パターン検出の際に、意味は類似しているが、連結構造が異なる構造を、全く別のパターンと判定してしまう、ということである。
【００１０】
連結構造とは、構造の節点と単語文字列及び有向枝の連結関係と、向きにのみ注目し、付属的な属性情報を省略した構造のことをいう。
【００１１】
上記第１の問題点が生じる理由は、従来のテキストマイニング装置は、連結構造が異なり、類似した意味を持つ構造を同一と判定する手段を具備していないためである。
【００１２】
属性値を用いた文構造を用いる際に、連結構造が異なり、類似した意味を持つ構造の差異の例として、
（Ｂ１）係り受けの向きの差異、
（Ｂ２）係り受けの順序の差異、
（Ｂ３）同義語の置換による差異、及び、
（Ｂ４）並列の構文構造と意味構造の差異
などが挙げられる。
【００１３】
図１３に、これらの連結構造による構造の差異の例を示す。また、属性値を用いない文構造を用いる際は、あらゆる意味の類似した構造の差異は、連結構造の差異で表現される。
【００１４】
図１３（ａ）に示す例では、意味の類似した「速いのは車種Ａ」と「車種Ａは速い」のの連結構造において、係り元と係り先が相違している。
【００１５】
図１３（ｂ）に示す例では、意味の類似した「速く安い車種Ａ」と「安く速い車種Ａ」の連結構造において、係り元の「速い」と「安い」の節点の順序関係が、相違している。
【００１６】
図１３（ｃ）に示す例では、意味の類似した「車種Ａは速い」と「車種Ａは高速だ」のそれぞれの連結構造において、係り先の「速い」と「高速」が相違している。
【００１７】
図１３（ｄ）に示す例では、「車種Ａと車種Ｂは速い」の構文構造と意味構造を表わしている。図３（ｄ）において、係り元「車種Ａ」が「車種Ｂ」に係り「車種Ｂ」が「速い」に係る連結構造と、係り元「車種Ａ」と「車種Ｂ」から係り先「速い」への有向枝を有する連結構造がある。
【００１８】
第２の問題点は、頻出パターン検出の際に異なる属性値を持つが、類似した意味を持つ構造を、全く別のパターンと判定してしまう、ということである。
【００１９】
その理由は、従来のテキストマイニング装置では、異なる属性値を持つ構造を、同一と判定することについて、何ら考慮されていないためである。
【００２０】
属性値を用いた文構造を用いる際に、属性値が異なり、類似した意味を持つ構造の差異の例として、付属語情報の差異、表層格の差異などが挙げられる。図１４に、これらの属性値による構造の差異の例を示す。
【００２１】

図１４（ａ）に示す例では、類似した意味を持つ「車種Ａは加速」と「車種Ａの加速」の連結構造において、有向枝の表層格が相違している。

【００２２】

図１４（ｂ）に示す例では、類似した意味を持つ「車種Ａは速い」と「車種Ａは速かった」の連結構造において、係り先の節点「速い」の付属語情報が相違している。

【００２３】

第３の問題点は、テキストマイニング装置の使用者（ユーザ）がどこまで類似した構造を同一な構造と判定して頻出パターンの検出を行うかを調整できない、ということである。

【００２４】

その理由は、従来のテキストマイニング装置では、使用者が頻出パターン検出の際にどのような構造を同一と判定するかを調整することについて、何ら考慮されていないためである。

【００２５】

したがって、本発明の目的は、類似した意味を持ち、連結構造の異なる構造を、同一のパターンと判定して頻出パターン等の検出を行うテキストマイニング装置及び方法並びにプログラムを提供することにある。

【００２６】

本発明の他の目的は、類似した意味を持ち属性値の異なる構造を同一な構造と判定して頻出パターン検出を行うかを調整できるテキストマイニング装置及び方法並びにプログラムを提供することにある。

【００２７】

本発明のさらに他の目的は、テキストマイニングの使用者がどこまで類似した構造を同一な構造と判定して頻出パターン検出を行うかを調整できるテキストマイニング装置及び方法並びにプログラムを提供することにある。

【課題を解決するための手段】

【００２８】

本願で開示される発明は、上記目的を達成するため、概略以下の構成とされる。

【００２９】

本発明の第１のアスペクト（側面）に係るテキストマイニング装置は、入力した文書から文構造を作成する手段と、前記文構造の部分構造に対して予め定められた所定の変換操作を行うことで、前記部分構造と意味の類似したパターンの類似構造を作成する手段と、前記意味の類似したパターンを同一パターンと判定してパターン検出を行う手段と、を備えている。

【００３０】

本発明において、前記類似構造を生成する手段は、前記文構造について並列変形を行う手段と、前記文構造の部分構造を生成する手段と、前記文書構造及び／又は部分構造の有向枝の無向枝化を行う手段と、同義語辞書を参照して前記文書構造及び／又は部分構造中の同義語の置換を行う手段と、前記文書構造及び／又は部分構造における順序木の無順序木化を行う手段と、を備え、前記類似構造を前記部分構造の同値類とする。同値類とは、構造の集合でその各要素を同一の構造として扱うものをいい、二つの同値類に一つでも、同一の要素が含まれる時には、その二つの同値類を同一の同値類と判定する。本発明によれば、生成された類似構造を生成元の文構造の同値類として扱い、頻出パターン検出を行う。

【００３１】

本発明の第２のアスペクト（側面）に係るテキストマイニング装置は、第１のアスペクトに係るテキストマイニング装置の構成に含まれる頻出パターン検出手段に代わり、構造中の属性値の差異を無視して、頻出パターンの検出を行う頻出類似パターン検出手段を備え、属性値の異なる類似した構造を同一な構造と判定して頻出パターンの検出を行う。本発明によれば、構造中の属性値が異なる類似した構造を同一と判定して頻出パターン検出

を行う。

【００３２】

本発明の第３のアスペクト（側面）に係るテキストマイニング装置は、テキストマイニングの対象となる文書の集まりを記憶する記憶部と、前記記憶部の前記文書を解析して文構造を取得する解析部と、使用者の入力から文構造の差異の種別ごとに同一構造と判定するか否かを指定する第１の指定項目を生成する類似構造生成調整部と、使用者の入力から属性値の差異の種別ごとに同一構造と判定するか否かを指定する第２の指定項目を生成し類似構造判定調整部と、前記類似構造生成調整部によって生成された第１の指定項目に従い、前記解析部で得られた文構造の部分構造に対して所定の変換操作を行い、前記部分構造と意味的に類似した類似構造を生成する類似構造生成部と、前記類似構造生成部によって生成された類似構造を生成元の部分構造の同値類として扱い、前記類似構造判定調整部の第２の指定項目に従い、属性値の差異を無視しながら、頻出パターンの検出を行う類似パターン検出部と、を備えている。本発明によれば、構造の同一性の判定を調整するための指定の入力を受け付ける。

【００３３】

本発明のさらに他のアスペクト（側面）に係る方法は、

入力した文書から文構造を作成する工程と、

前記文構造の部分構造に対する所定の変換操作を行うことで、前記部分構造と意味の類似したパターンの類似構造を作成する工程と、

前記意味の類似したパターンを同一パターンと判定してパターン検出を行う工程と、を含む。

【００３４】

本発明のさらに他のアスペクト（側面）に係る方法は、テキストマイニングの対象となるテキストの集まりを記憶する記憶部のテキストを解析して文構造を取得する工程と、

前記文構造の部分構造に対して意味的に類似しパターンの類似構造を生成する工程と、

生成された類似構造を生成元の部分構造の同値類として扱い、属性値の差異を無視しながらパターンの検出を行う工程と、を含む。

【００３５】

本発明のさらに他のアスペクト（側面）に係る方法は、テキストマイニングの対象となるテキストの集まりを記憶する記憶部のテキストを解析して文構造を取得する工程と、

入力装置から入力された使用者の入力情報から、文構造（連結構造）の差異の種別ごとに同一構造と判定するか否かを指定する第１の指定項目と、属性値の差異の種別ごとに同一構造と判定するか否かを指定する第２の指定項目を生成するステップと、

文構造（連結構造）の差異の種別ごとに同一構造と判定するか否かを指定する第１の指定項目に従い、前記文構造の部分構造に対して意味的に類似した構造を生成する工程と、

生成された類似構造を生成元の部分構造の同値類として扱い、属性値の差異の種別ごとに同一構造と判定するか否かを指定する第２の指定項目に従い、属性値の差異を無視しながら頻出パターンの検出を行う工程と、を含む。

【００３６】

本発明のさらに他のアスペクト（側面）に係るプログラムは、テキストマイニング装置を構成するコンピュータに、

テキストマイニングの対象となるテキストの集まりを記憶する記憶部の前記テキストを解析して文構造を取得する処理と、

前記処理で解析して得られた文構造の部分構造に対して、意味的に類似した構造を生成する処理と、

生成された類似構造を、生成元の部分構造の同値類として扱い、頻出パターンの検出を行う処理と、

を実行させるプログラムよりなる。

【発明の効果】

【００３７】

本発明によれば、連結構造は異なるが類似した意味を持つ構造を同一の構造と判定して頻出パターンを検出することができる。本発明によれば、属性値を持たない構造の集合に対して類似構造を同一と判定して頻出パターンの検出を行うことができる。

【００３８】

その理由は、本発明においては、生成した類似構造を元の構造の同値類として扱い、頻出パターン検出を行う構成としたためである。本発明によれば、属性値を持つ構造の集合に対しても類似構造を同一と判定して頻出パターンの検出を行うことができる。

【００３９】

また、本発明によれば、類似した意味を持つが異なる属性値を持つ構造を同一の構造と判定して頻出パターンを検出することができる。

【００４０】

その理由は、本発明においては、頻出類似パターン検出手段が属性値の差異を無視して頻出パターン検出を行うためである。

【００４１】

さらに本発明によれば、テキストマイニング装置の使用者がどこまで類似した構造を同一な構造と判定して頻出パターン検出を行うかを調整することができる。

【００４２】

その理由は、本発明においては、類似構造生成調整手段と類似構造判定調整手段が使用者からの入力に基づき、どこまで類似した構造を同一な構造と判定するかの調整を行う構成としたためである。

【発明を実施するための最良の形態】

【００４３】

次に、発明を実施するための最良の形態について図面を参照して詳細に説明する。

【００４４】

図１を参照すると、本発明を実施するための最良の一形態（第１の実施の形態）は、情報を記憶する記憶装置１と、プログラム制御により動作するデータ処理装置２と、検出されたパターンを出力する出力装置３と、を備えている。記憶装置１はテキストデータベース（ＤＢ）１１を含む。テキストＤＢ１１は、テキストマイニングの対象となるテキストの集合を記憶している。

【００４５】

データ処理装置２は、言語解析手段２１と、類似構造生成手段２２と、頻出パターン検出手段２３を含む。これらの手段はそれぞれ概略つぎのように動作する。

【００４６】

言語解析手段２１は、テキストＤＢ１１からテキスト集合を読み込み、集合中の各テキストを解析して文構造を得る。

【００４７】

類似構造生成手段２２は、言語解析手段２１から送られた文構造の集合中の各文構造の全ての部分構造を抽出し、各部分構造の全ての類似構造を生成して類似構造と生成元の部分構造を同値類とする。

【００４８】

頻出パターン検出手段２３は、類似構造生成手段２２から送られた部分構造の同値類の集合から頻出するパターンを検出し、出力装置３へ送る。

【００４９】

図２は、本実施形態の動作を説明するための流れ図である。次に、図１及び図２を参照して、本発明の第１の実施形態の動作について詳細に説明する。

【００５０】

まず、言語解析手段２１が、テキストＤＢ１１から、テキスト集合を読み込む。言語解析手段２１は、テキスト集合中の各テキストに対し解析を行い、解析結果として、文構造を生成し、類似構造生成手段２２に送る（図２のステップＡ１）。

【００５１】

次に、類似構造生成手段２２は、与えられた文構造の集合中の部分構造の全ての類似構造を生成し、類似構造を生成元の部分構造の同値類とし、同値類の集合を頻出パターン検出手段２３に送る（図２のステップＡ２）。

【００５２】

さらに、頻出パターン検出手段２３は、与えられた部分構造の同値類から、頻出パターンの検出を行う（図２のステップＡ３）。

【００５３】

頻出パターン検出手段２３は、検出した頻出パターンを出力装置３に出力する（図２のステップＡ４）。

【００５４】

図３は、図２のステップＡ２における、類似構造生成手段２２の動作の詳細なフローチャートを示す図である。

【００５５】

図３を参照すると、類似構造生成手段２２は、まず並列構文の構文構造と意味構造の違いに対応するための「並列の変形」を行う（図３のステップＡ２－１）。

【００５６】

次に、文構造全体だけではなく部分構造からもパターン検出を行うための「部分構造の生成」を行う（図３のステップＡ２－２）。

【００５７】

次に、係り受けの向きの差異に対応するための「有向枝の無向枝化」を行う（図３のステップＡ２－３）。

【００５８】

次に、同義語の差異に対応するための「同義語の置換」を行う（図３のステップＡ２－４）。

【００５９】

その係り受けの順序の違いに対応するための「順序木の無順序木化」を行う（図３のステップＡ２－５）。

【００６０】

最後に、類似構造を生成元の部分構造の同値類の要素とすることで、「同値類の生成」を行う（図３のステップＡ２－６）。

【００６１】

次に、本発明の第１の実施の形態の作用効果について説明する。

【００６２】

本発明の第１の実施の形態では、類似構造生成手段２２が生成した類似構造を、元の構造の同値類として扱い、頻出パターン検出を行うように構成されている。このため、連結構造は異なるが、類似した意味を持つ構造を、同一の構造と判定して、頻出パターンを検出できる。

【００６３】

次に、本発明を実施するための最良の別の形態（第２の実施の形態）について図面を参照して詳細に説明する。

【００６４】

図４を参照すると、本発明の第２の実施の形態においては、データ処理装置４が、図１に示された第１の実施の形態におけるデータ処理装置２の頻出パターン検出手段２３に代えて、頻出類似パターン検出手段２４を備えている。言語解析手段２１、類似構造生成手段２２は、前記第１の実施の形態のものと同一である。

【００６５】

本発明の第２の実施の形態において、頻出類似パターン検出手段２４は、類似構造生成手段２２から送られた部分構造の同値類の集合から、属性値の相違を無視しながら、頻出パターンの検出を行い、検出した頻出パターンを出力装置３に送る。

【００６６】

図５は、本発明の第２の実施形態の動作を説明するための流れ図である。次に、図４及び図５を参照して、本発明の第２の実施形態の動作について詳細に説明する。本発明の第２の実施形態においては、図２のステップＡ３のかわりに、ステップＢ３が実行される。図５のステップＡ１、Ａ２、Ａ４で示される処理は、前記第１の実施の形態の処理と同一であるため、説明は省略する。

【００６７】

前記第１の発明の実施の形態では、頻出パターン検出手段２３は、連結構造が同一でも属性値の異なる構造は同一と判定せずに、頻出パターンの検出を行っていた。

【００６８】

本実施の形態では、頻出類似パターン検出手段２４は、類似構造生成手段２２から与えられた同値類の集合を、連結構造が同一で属性値の異なる構造も同一な構造と判定しながら頻出パターンの検出を行い、検出された頻出パターンを出力装置３に送る（図５のステップＢ３）。

【００６９】

次に、本発明の第２の実施形態の作用効果について説明する。

【００７０】

本発明の第２の実施の形態では、頻出類似パターン検出手段２４が、連結構造は同一で属性値の異なる構造も同一な構造と判定しながら頻出パターンの検出を行うように構成されている。このため、意味は類似しているが属性値の異なる構造も同一な構造と判定して頻出パターンの検出を行うことができる。

【００７１】

次に、本発明の第３の発明を実施するための最良の形態について図面を参照して詳細に説明する。

【００７２】

図６を参照すると、本発明の第３の実施の形態は、図４に示した前記第２の実施の形態の構成と比較し、入力装置６を備え、データ処理装置５が、類似構造生成調整手段２５及び類似構造判定調整手段２６を備えている。

【００７３】

入力装置６は、使用者から、

・文構造の差異の種別ごとに同一構造と判定するか否かを指定するための入力と、

・属性値の種別ごとに値の差異を無視するか否かを指定するための入力と、

を受け付け、それぞれを、類似構造生成調整手段２５と類似構造判定調整手段２６に送る。

【００７４】

入力装置６で受け付ける指定の入力の例としては、

・「使用者から文構造の差異の種別ごとに同一構造と判定するか否かと属性値の種別ごとに値の差異を無視するか否かについての指定項目」、

・「頻出パターン検出の際に同一パターンを持っていると判定しない文の例」、

・「頻出パターン検出の際に同一パターンを持っていると判定する文の例」

などが挙げられる。

【００７５】

類似構造生成調整手段２５は、入力装置６から与えられた指定から、連結構造の差異の種別ごとに同一構造と判定するか否かを決定し、その指定項目を、類似構造生成手段２２に送る。

【００７６】

また、類似構造判定調整手段２６は、入力装置６から与えられた指定から、属性値の種別ごとに値の差異を無視するか否かを決定し、その指定項目を、頻出類似パターン検出手段２４に送る。

【００７７】

類似構造生成手段２２は、類似構造生成調整手段２５からの指定に従って、言語解析手

段２１より与えられた集合中の各構造の部分構造について、該部分構造の類似構造の生成を行い、生成された各類似構造を、それぞれの生成元の部分構造の同値類とする。

【００７８】

頻出類似パターン検出手段２４は、類似構造判定調整手段２６からの指定に従って、属性値の差異の無視を行いながら、類似構造生成手段２２より与えられた同値類の集合から頻出パターンの検出を行う。

【００７９】

図７は、本発明の第３の実施の形態の動作を説明するための流れ図である。次に、図６及び図７のフローチャートを参照して本発明の第３の実施の形態の動作について詳細に説明する。

【００８０】

まず、言語解析手段２１がテキストＤＢ１１からテキスト集合を読み込む。

【００８１】

言語解析手段２１は、テキスト集合中の各テキストに対して解析を行い、解析結果として文構造を生成し、類似構造生成手段２２に送る（図７のステップＡ１）。図７のステップＡ１における言語解析手段２１の動作は、前記第１の実施の形態の言語解析手段２１と同一である。

【００８２】

次に、入力装置６が、使用者から文構造の差異の種別ごとに同一構造と判定するか否かを指定するための入力と、属性値の種別ごとに値の差異を無視するか否かを指定するための入力とを受け付け、それぞれ類似構造生成調整手段２５と類似構造判定調整手段２６に送る（図７のステップＣ１）。

【００８３】

類似構造生成調整手段２５は、入力装置６から与えられた指定から、文構造の差異の種別ごとに同一構造と判定するか否かの指定項目を生成し、類似構造生成手段２２に送る。また、類似構造判定調整手段２６は、入力装置６から与えられた指定から属性値の種別ごとに値の差異を無視するか否かの指定項目を生成し頻出類似パターン検出手段２４に送る（図７のステップＣ２）。

【００８４】

類似構造生成手段２２は、類似構造生成調整手段２５からの指定に従って、言語解析手段２１より与えられた集合中の各文構造の部分構造の類似構造の生成を行い、生成された各類似構造を、それぞれの生成元の部分構造の同値類とし、同値類の集合を頻出類似パターン検出手段２４に送る（図７のステップＣ３）。

【００８５】

頻出類似パターン検出手段２４は、類似構造判定調整手段２６からの指定に従って属性値の無視を行いながら、類似構造生成手段２２より与えられた同値類の集合から頻出パターンの検出を行う（図７のステップＣ４）。

【００８６】

最後に、頻出類似パターン検出手段２４は、検出した頻出パターンを出力装置３に出力する（図７のステップＡ４）。

【００８７】

図８は、図７のステップＣ３における、類似構造生成手段２２の動作の詳細なフローチャートである。

【００８８】

図８を参照すると、類似構造生成手段２２は、

ステップＣ３－１の判定において、並列の変形が指定されている場合、並列の変形（図８のステップＡ２－１）を行い、部分構造の生成（図８のステップＡ２－２）を行い、並列の変形が指定されていない場合、ステップＡ２－２へ分岐する。並列の変形、部分構造の生成は、図３のステップＡ２－１、Ａ２－２と同一である。

【００８９】

ステップＣ３－２の判定において、有向枝の無向枝化が指定されている場合、有向枝の無向枝化（図８のステップＡ２－３）を行い、指定されていない場合、ステップＣ３－３に分岐する。有向枝の無向枝化は、図３のステップＡ２－３と同一である。

【００９０】

ステップＣ３－３の判定において、同義語の置換が指定されている場合、同義語の置換（図８のステップＡ２－４）を行い、同義語の置換が指定されていない場合、ステップＣ３－４に進む。同義語の置換は、図３のステップＡ２－４と同一である。

【００９１】

ステップＣ３－３の判定において、順序木の無順序木化が指定されている場合、順序木の無順序木化（図８のステップＡ２－５）を行い、指定されていない場合、ステップＡ２－６に分岐する。

【００９２】

ステップＡ２－６では、同値類を生成する。順序木の無順序木化、同値類を生成は、図３のステップＡ２－５、Ａ２－６と同一である。

【００９３】

このように、本発明の第３の実施の形態では、並列の変形（図８のステップＡ２－１）、有向枝の無向枝化（図８のステップＡ２－３）、同義語の置換（図８のステップＡ２－４）、及び、順序木の無順序木化（図８のステップＡ２－５）が、類似構造生成調整手段２５から与えられた指定により、実行の有無が制御される点で、図３に示した前記第１の実施の形態の類似構造生成手段２２と相違している。

【００９４】

使用者は、出力されたパターンを参照して、ステップＣ１に戻りどこまで類似した構造を同一と判定するかを指定するための入力を再度行ったうえで本発明に頻出パターン検出を再度行わせることができる。

【００９５】

次に、本発明の第３の実施の形態の作用効果について説明する。

【００９６】

本発明の第３の実施の形態では、類似構造生成調整手段と類似構造判定調整手段が使用者からの指定に基づきどこまで類似した構造を同一な構造と判定するかの調整を行うように構成されているため、使用者がどこまで類似した構造を同一な構造と判定して頻出パターン検出を行うかを調整できる。

【００９７】

次に、本発明の第４の実施の形態について図面を参照して詳細に説明する。

【００９８】

図９を参照すると、本発明の第４の実施の形態は、前記した第１、第２、第３の実施の形態をプログラムにより構成した場合に、そのプログラムにより動作されるコンピュータの構成を示す図である。

【００９９】

テキストマイニング用プログラム７は、データ処理装置８に読み込まれ、データ処理装置８の動作を制御する。データ処理装置８はテキストマイニング用プログラム７の制御により以下の処理、すなわち第１、第２及び第３の実施の形態におけるデータ処理装置２、４及び５による処理と同一の処理を実行する。

【実施例】

【０１００】

次に、本発明を具体的な実施例を即して詳細に説明する。

【０１０１】

まず、本発明の第１の実施例を、図面を参照して説明する。本発明の第１の実施例は、前記第１の実施の形態の一具体例である。

【０１０２】

本発明の第１の実施例においては、図１のデータ処理装置２をパーソナル・コンピュー

タで構成し、記憶装置１を、磁気ディスク記憶装置で構成し、出力装置３としてディスプレイを備えている。

【０１０３】

パーソナル・コンピュータ２は、言語解析手段２１、類似構造生成手段２２、頻出パターン検出手段２３として機能する中央演算装置（ＣＰＵ）を有している。磁気ディスク記憶装置には、テキストＤＢ１１としてテキスト集合が記憶されている。

【０１０４】

図１５は、テキスト集合の内容を示す図である。

【０１０５】

言語解析手段２１は、テキストＤＢ１１中の図１５に示されるテキスト集合の各テキストに対して言語解析を行い、各テキストの文構造を得る（図２のステップＡ１）。

【０１０６】

図１６（ａ）乃至（ｃ）に、言語解析手段２１で得られる文構造（文１～文３）を示す。

【０１０７】

次に、類似構造生成手段２２は、図１６（ａ）乃至（ｃ）に示される各文構造の部分構造の全ての類似構造を生成し、生成された類似構造を、生成元の部分構造の同値類とする（図２のステップＡ２）。

【０１０８】

本実施例では、図１６（ｂ）に示される文２（「速く安い車種Ａ」）の文構造から、部分構造の同値類を生成する様子を例にとって説明する。この例は、図１８乃至２１に示されている。

【０１０９】

まず、図１８に示すように、並列構造の変形を行う（図３のステップＡ２－１）。図１８に示す例においては、部分構造２ａ－０において、並列関係にある「速い」と「安い」の接続関係を変形し、類似構造２ａ－１を生成している。

【０１１０】

次に、図１９に示すように、部分構造の生成を行う（図３のステップＡ２－２）。図１９に示す例においては、部分構造２ａ－０から、２単語の関係を表す部分構造２ｃ－０及び２ｇ－０と、１単語の部分構造２ｄ－０、２ｅ－０及び２ｆ－０を生成する。

【０１１１】

また、類似構造２ａ－１から、部分構造２ａ－０に含まれない２単語の関係を表す部分構造２ｂ－０を生成する。

【０１１２】

なお、部分構造２ａ－０と類似構造２ａ－１の両方から生成される構造は１つにまとめて扱う。

【０１１３】

また、ここで部分構造を生成するのに用いた部分構造２ａ－０、及び類似構造２ａ－１も、今後の類似構造生成において、部分構造及び類似構造として扱う。

【０１１４】

次に、図２０に示すように、有向枝の無向枝化を行う（図３のステップＡ２－３）。この例においては、ステップＡ２－２において生成した部分構造の全ての有向枝を無向枝化して、新たな類似構造を生成している。図２０(ａ)に示すように、例えば部分構造２ａ－０の有向枝を無向枝化して、類似構造２ａ－２が生成される。なお、１単語からなり有向枝を持たない部分構造２ｄ－０、２ｅ－０及び２ｆ－０は、ステップＡ２－３では変形が行われないため、図２０では省略されている。

【０１１５】

次に、同義語の置換が行われる（図３のステップＡ２－４）。本実施例における「同義語の置換」では、ユーザによりあらかじめ与えられた同義語辞書に定義された被置換語を代表語に置き換えるものとする。

【０１１６】
また、本実施例に用いる同義語辞書は、図１７に示されるように、被置換語「高速」を代表語「速い」に置き換える１つの辞書項目のみが登録された同義語辞書が指定されたものとしている。
【０１１７】
ここで、図２０に示された、この時点で生成された部分構造及び類似構造には、被置換語「高速」が含まれないため、ステップＡ２－４では、変形が発生しない。そのため、ここではステップＡ２－４による変形の図を省略している。
【０１１８】
次に、順序木の無順序木化が行われる（図３のステップＡ２－５）。ここでは、文構造の木構造において、兄弟関係にある単語を５０音順にソートすることによって、順序木の無順序木化を行う。
【０１１９】
なお、順序木の無順序木化を行うための他の方法として、
・兄弟関係にある単語を５０音順以外の一定の法則に従いソートする方法や、
・ソートを行わずに頻出類似パターン検出時に兄弟関係にある単語の順序だけが異なる木を同一と判定する方法を用いてもよい。
【０１２０】
図２０に示された、この時点で、生成された部分構造及び類似構造では、類似構造２ａ－１及び２ａ－３を除いた部分構造及び類似構造では兄弟関係になっている単語が存在せず、類似構造２ａ－１及び２ａ－３では、既に兄弟関係にある単語が５０音順に並んでいるため、実質的に変形が発生しない。そのため、ここではステップＡ２－５による変形の図を省略している。
【０１２１】
最後に、類似構造を生成元の部分構造の同値類とすることで、同値類の生成を行う（図３のステップＡ２－６）。
【０１２２】
図２０に示された部分構造及び類似構造の集合において、各類似構造を生成元の部分構造の同値類することで生成される同値類を図２１に示す。部分構造２ａ－０と、部分構造２ａ－０の有向枝を無向枝化することで生成された類似構造２ａ－２と、部分構造２ａ－０を並列変形した類似構造２ａ－１と、類似構造２ａ－１の有向枝を無向枝化することで生成された類似構造２ａ－３とは同値類２ａを構成している。部分構造２ｂ－０と、部分構造２ｂ－０の有向枝を無向枝化することで生成された類似構造２ｂ－１は、同値類２ｂを構成している。部分構造２ｃ－０と、部分構造２ｃ－０の有向枝を無向枝化することで生成された類似構造２ｃ－１は、同値類２ｃを構成している。部分構造２ｇ－０と、部分構造２ｇ－０の有向枝を無向枝化することで生成された類似構造２ｇ－１は、同値類２ｇを構成している。部分構造２ｄ－０、２ｅ－０、２ｆ－０は、類似構造と部分構造は同一である。
【０１２３】
図１８乃至図２１に示したように、本実施例において、文２の文構造から、類似構造生成手段２２が同値類を生成する例においては、同義語の置換（図３のステップＡ２－４）及び順序木の無順序木化（図３のステップＡ２－５）で変形は行われない。
【０１２４】
図２２に示すように、文３の文構造の１つの部分構造に対して、類似構造生成手段２２が行う変形を用いて、同義語の置換（図３のステップＡ２－４）及び順序木の無順序木化（図３のステップＡ２－５）で発生する変形の例を説明する。
【０１２５】
まず文３の文構造を表す部分構造３ａ－０に対して並列の変形（図３のステップＡ２－１）が行われる。ここでは、部分構造３ａ－０が並列の構造を含まず変形が行われないため、図２２には、並列の変形による結果の構造は含まれない。

【０１２６】
次に、部分構造３ａ－０から部分構造の生成（図３のステップＡ２－２）が行われる。ここでは、部分構造３ａ－０に行われる構造変形にのみ注目して説明するため、部分構造３ａ－０から、他の部分構造を生成する処理である部分構造の生成は省略する。
【０１２７】
次に、部分構造３ａ－０に対して、有向枝の無向枝化（図３のステップＡ２－３）が行われる。図２２（ａ）の部分構造３ａ－０の「安い」から、「車種Ａ」への有向枝と「高速」から「車種Ａ」への有向枝が無向枝化され、図２２（ｂ）の類似構造３ａ－１が生成される。
【０１２８】
次に、類似構造３ａ－１に対して、同義語の置換（図３のステップＡ２－４）が行われる。ここでは、図１７に示される同義語辞書を用いているため、被置換語「高速」が代表語「速い」に置き換えられる。類似構造３ａ－１に含まれる被置換語「高速」も代表語「速い」に置き換えられ、図２２（ｃ）の類似構造に変形される。
【０１２９】
次に、図２２（ｃ）の類似構造３ａ－１に対して、順序木の無順序木化（図３のステップＡ２－５）が行われる。ここでは、兄弟関係にある単語を、５０音順にソートすることで順序木の無順序木化が行われるため、類似構造３ａ－１において兄弟関係にある「安い」と「速い」の順序を入れ替え、５０音順にソートにされ、図２２（ｄ）の類似構造に変換される。
【０１３０】
このようにして生成された類似構造に対して同値類の生成（図３のステップＡ２－６）が行われるが、本実施例では、部分構造３ａ－０から生成される一つの類似構造３ａ－１に行われる変形のみに注目して説明しているため省略する。
【０１３１】
このようにして、類似構造生成手段２２が、部分構造と類似構造及び同値類の生成を行うことで、本実施例では、図１６の文１の文構造から、図２３に示すような同値類が生成される。図１６の文２の文構造から、図２４に示すような同値類が生成される。また図１６の文３の文構造から、図２５に示すような同値類が生成される。
【０１３２】
ただし、本来は、図２２における変形の途中経過（図２２（ｂ）、図２２（ｃ）の類似構造３ａ－１）のように、形の違う類似構造も生成されているが、説明を分かりやすくするため、頻出パターンの検出に用いられない構造は、図２３乃至図２５の同値類からは省略している。
【０１３３】
次に、頻出パターン検出手段２３は、図２３乃至図２５に示される同値類の集合から頻出パターン（頻出する同値類）の検出を行う（図２のステップＡ３）。
【０１３４】
この際、頻出パターン検出手段２３は、要素の少なくとも一つが同一である同値類は、同一と判定して、頻出パターンの検出を行う。
【０１３５】
例えば、本実施例においては、図２３の同値類１ｃの要素である類似構造１ｃ－１と、図２４の同値類２ｂの要素である類似構造２ｂ－１は、どちらも「車種Ａ」と「速い」が無向枝で連結された構造で、属性値の差分もないため、同一の構造である。
【０１３６】
従って、頻出パターン検出手段２３は、図２３の同値類１ｃと図２４の同値類２ｂを同一と判定する。
【０１３７】
図２３乃至図２５を参照すると、
「類似構造１ｃ－１、類似構造２ｂ－１と、類似構造３ｃ－１」、

「部分構造１ｄ－０、部分構造２ｄ－０と、類似構造３ｅ－１」、
「部分構造１ｅ－０、部分構造２ｆ－０と、部分構造３ｆ－０」、
「部分構造１ｆ－０と部分構造２ｅ－０」
がそれぞれ同一の構造となっている。

【０１３８】

「要素の少なくとも一つが同一である同値類は同一と判定する」という同値類の性質により、図２３乃至図２５に示される同値類のうち、
「同値類１ｃ、２ｂ、及び、３ｃ」、
「同値類１ｄ、２ｄ、及び、３ｅ」、
「同値類１ｅ、２ｆ、及び、３ｆ」、
「同値類１ｆ、及び、２ｅ」
がそれぞれ同一の同値類と判定される。

【０１３９】

本実施例では、３回以上出現する同値類を頻出パターンとする。なお、どのような出現回数の同値類を頻出パターンとして検出するかは、使用者がテキストマイニングを実行する前に決定することができる。

【０１４０】

この場合、
「同値類１ｃ、２ｂ、及び、３ｃ」、
「同値類１ｄ、２ｄ、及び、３ｅ」、
「同値類１ｅ、２ｆ、及び、３ｆ」
が頻出パターンとして検出される。

【０１４１】

最後に、そのようにして抽出された頻出パターンを表す構造を出力装置３に表示する（図２のステップＡ４）。

【０１４２】

図２６は、本実施例において、出力装置３が出力する頻出パターンの表現の一例を示す図である。本実施例では、頻出パターンを表す同値類の要素である類似構造を、頻出パターンの表現として用いている。

【０１４３】

このようにして、類似構造を生成し、同値類を生成して頻出パターンの検出を行うことで、「部分構造１ｃ－０（図２３）、部分構造２ｂ－０（図２４）、及び、部分構造３ｃ－０（図２５）」のように類似した意味を持つが、連結構造の異なる部分構造を同一と判定し、頻出パターンとして検出することができる。

【０１４４】

次に、本発明の第２の実施例を、図面を参照して説明する。本実施例は、前記第２の実施の形態に対応するものである。

【０１４５】

本発明の第２の実施例は、データ処理装置４をパーソナル・コンピュータで構成し、記憶装置１を磁気ディスク記憶装置で構成し、出力装置３としてディスプレイを備えている。

【０１４６】

パーソナル・コンピュータ４は、言語解析手段２１、類似構造生成手段２２、頻出類似パターン検出手段２４として機能する中央演算装置（ＣＰＵ）を有し、磁気ディスク記憶装置には、テキストＤＢ１１としてテキスト集合が記憶されている。テキスト集合としては、前記第１の実施例と同様、図１５に示した文１乃至文３を使用する。

【０１４７】

言語解析手段２１は、テキストＤＢ１１中の図１５に示されるテキスト集合の各テキストに対して、言語解析を行い、各テキストの文構造を得る（図５のステップＡ１）。ここで得られる文構造は、前記第１の実施例と同様、図１６のようになる。

【０１４８】
次に、類似構造生成手段２２は、図１６に示される各文構造の部分構造の全ての類似構造を生成し、生成された類似構造を生成元の部分構造の同値類とする（図５のステップＡ２）。ここで得られる同値類は、前記第１の実施例と同様、図２３乃至図２５のようになる。

【０１４９】
次に、頻出類似パターン検出手段２４は、図２３乃至図２５に示される同値類の集合から、属性値の差異を無視しながら頻出パターン（頻出する同値類）の検出を行う（図５のステップＢ３）。

【０１５０】
頻出類似パターン検出手段２４は、要素の少なくとも一つが同一である同値類は同一と判定して、頻出パターンの検出を行う。ただし、本実施例の頻出類似パターン検出手段２４は、表層格や付属語情報などの属性値の差異を無視して、類似構造の同一性の判定を行っており、この点で、前記第１の実施例の頻出パターン検出手段２３と相違している。

【０１５１】
例えば、図２３の類似構造１ａ－１と図２４の類似構造２ａ－３は、どちらも、「車種Ａ」と「速い」及び「安い」が無向枝で連結された構造であるが、表層格が異なるため、前記第１の実施例の頻出パターン検出手段２３では、同一と判定されない。一方、本実施例の頻出類似パターン検出手段２４では、同一と判定される。

【０１５２】
本実施例においては、図２３乃至図２５を参照すると、
「類似構造１ａ－１、類似構造２ａ－３、及び、類似構造３ａ－１」、
「類似構造１ｂ－１、類似構造２ｃ－１と、類似構造３ｂ－１」、
「類似構造１ｃ－１、類似構造２ｂ－１、及び、類似構造３ｃ－１」、
「部分構造１ｄ－０、部分構造２ｄ－０、及び、類似構造３ｅ－１」、
「部分構造１ｅ－０、部分構造２ｆ－０、及び、部分構造３ｆ－０」、
「部分構造１ｆ－０、部分構造２ｅ－０、及び、部分構造３ｄ－０」
がそれぞれ頻出類似パターン検出手段２４に同一の構造と判定される。

【０１５３】
頻出類似パターン検出手段２４は、要素の少なくとも一つが同一である同値類は同一と判定するため、
「同値類１ａ、２ａ、及び、３ａ」、
「同値類１ｂ、２ｃ、及び、３ｂ」、
「同値類１ｃ、２ｂ、及び、３ｃ」、
「同値類１ｄ、２ｄ、及び、３ｅ」、
「同値類１ｅ、２ｆ、及び、３ｆ」、
「同値類１ｆ、２ｅ、及び、３ｄ」
をそれぞれ同一の同値類と判定する。

【０１５４】
本実施例では、前記第１の実施例と同様に、３回以上出現する同値類を頻出パターンとする。

【０１５５】
この場合、
「同値類１ａ、２ａ、及び、３ａ」、
「同値類１ｂ、２ｃ、及び、３ｂ」、
「同値類１ｃ、２ｂ、及び、３ｃ」、
「同値類１ｄ、２ｄ、及び、３ｅ」、
「同値類１ｅ、２ｆ、及び、３ｆ」、
「同値類１ｆ、２ｅ、及び、３ｄ」
が頻出パターンとして検出される。

【０１５６】
最後に、そのようにして抽出された頻出パターンを表す構造を、出力装置３に表示する（図５のステップＡ４）。
【０１５７】
本実施例において、出力装置３が出力する頻出パターンの表現は図２７のようになる。本実施例では、前記第１の実施例と同様に、頻出パターンを表す同値類の要素である類似構造を、頻出パターンの表現として用いている。
【０１５８】
このようにして、属性値の差異を無視して頻出パターンの検出を行うことで、
「部分構造１ｂ－０（図２３）、部分構造２ｃ－０（図２４）と部分構造３ｂ－０（図２５）」、
「部分構造１ｆ－０（図２３）、部分構造２ｅ－０（図２４）と部分構造３ｆ－０（図２５）」
のように、類似した意味を持つが属性値の異なる部分構造を同一と判定し、頻出パターンとして、検出を行うことができる。
【０１５９】
次に、本発明の第３の実施例を、図面を参照して説明する。本実施例は、本発明の第３の実施の形態に対応するものである。
【０１６０】
本発明の第３の実施例は、データ処理装置５をパーソナル・コンピュータで構成し、記憶装置１を磁気ディスク記憶装置で構成し、出力装置３としてディスプレイを、入力装置６としてキーボードを備えている。
【０１６１】
パーソナル・コンピュータ５は、言語解析手段２１、類似構造生成手段２２、頻出類似パターン検出手段２４、類似構造生成調整手段２５、類似構造判定調整手段２６として機能する中央演算装置（ＣＰＵ）を有している。磁気ディスク記憶装置には、テキストＤＢ１１としてテキスト集合が記憶されている。テキスト集合としては、前記第１、第２の実施例と同様、図１５に示した文が用いられる。
【０１６２】
言語解析装置２１は、テキストＤＢ１１中の図１５に示されるテキスト集合の各テキストに対して、言語解析を行い、各テキストの文構造を得る（図７のステップＡ１）。ここで得られる文構造は、前記第１、第２の実施例と同じく、図１６のようになる。
【０１６３】
次に、使用者は、入力装置６を用いて、
・文構造の差異の種別ごとに同一構造と判定するか否かを指定するための入力と、
・属性値の種別ごとに値の差異を無視するか否かを指定するための入力
を行う（図７のステップＣ１）。
【０１６４】
本実施例では、
「連結構造の差異については、係り受けの向きの差異と係り受けの順序の差異は同一と判定し、同義語の置換による差異は同一と判定しない。属性値の差異については、付属語情報の差異と表層格の差異は同一と判定する」という入力を行ったとする。
【０１６５】
入力装置６は、使用者から受け付けた入力を、類似構造生成調整手段２５と類似構造判定調整手段２６に送る。
【０１６６】
次に、類似構造生成調整手段２５は、入力装置６から使用者の指定を受け取り、類似構造生成手段２２の動作を制御する（図７のステップＣ２）。
【０１６７】
本実施例においては、類似構造生成調整手段２５は、入力装置６から、

「連結構造の差異については、係り受けの向きの差異と係り受けの順序の差異は同一と判定し、同義語の置換による差異は同一と判定しない。属性値の差異については、付属語情報の差異と表層格の差異は同一と判定する」

という指定を受け取り、類似構造生成手段２２が、文構造の部分構造から類似構造を生成する際の変形処理において、並列構造の変形（図８のステップＡ２－１）、有向枝の無向枝化（図８のステップＡ２－３）及び順序木の無順序木化（図８のステップＡ２－５）は行われるが、同義語の置換（図８のステップＡ２－４）はスキップされるように制御する。

【０１６８】

一方、類似構造判定調整手段２６は、入力装置６から使用者の入力を受け取り、頻出類似パターン検出手段２４の動作を制御する（図７のステップＣ２）。

【０１６９】

本実施例においては、類似構造生成調整手段２６は、入力装置６から、「連結構造の差異については、係り受けの向きの差異と係り受けの順序の差異は同一と判定し、同義語の置換による差異は同一と判定しない。属性値の差異については、付属語情報の差異と表層格の差異については同一と判定する」という指定を受け取り、頻出類似パターン検出手段２４が類似構造の同一性判定の処理を、表層格の差異を無視し、付属語情報の差異も無視して行うように制御する。

【０１７０】

次に、類似構造生成手段２２は、図１６に示される各文構造の部分構造についてステップＣ２で生成した指定項目に従い、同義語の置換（図８のステップＡ２－４）を飛ばして類似構造を生成し、生成された類似構造を生成元の部分構造の同値類とする（図７のステップＣ３）。

【０１７１】

以下、図１６に示される文３の文構造の１つの部分構造に対して、類似構造生成手段２２が行う変形を例にとって説明する。図２８に、一例を示す。

【０１７２】

まず、文３の文構造を表す部分構造３ａ－０に対して、並列の変形（図８のステップＡ２－１）が行われる。ただし、図２８に示す例では、部分構造３ａ－０が並列の構造を含まず変形が行われないため、図２８には、並列の変形による結果の構造は含まれない。

【０１７３】

次に、部分構造３ａ－０から部分構造の生成（図８のステップＡ２－２）が行われるが、ここでは、部分構造３ａ－０に行われる構造変形にのみ注目して説明するため、部分構造３ａ－０から他の部分構造を生成する処理である部分構造の生成は省略する。

【０１７４】

次に、部分構造３ａ－０に対して有向枝の無向枝化（図８のステップＡ２－３）が行われる。部分構造３ａ－０の「安い」から「車種Ａ」への有向枝と、「高速」から「車種Ａ」への有向枝が無向枝化され、図２８（ｂ）の類似構造３ａ－２が生成される。

【０１７５】

同義語の置換（図８のステップＡ２－４）は、類似構造生成調整手段２５より与えられた指定により、ステップＣ３－３の判定でスキップされるため、実行されない。

【０１７６】

次に、類似構造３ａ－２に対して、順序木の無順序木化（図８のステップＡ２－５）が行われる。ここでは、兄弟関係にある単語を５０音順にソートすることで、順序木の無順序木化が行われる。図２８（ｂ）の類似構造３ａ－２において、兄弟関係にある「安い」と「高速」の順序を入れ替えて、５０音順にソートされ、図２８（ｃ）の構造に変換される。

【０１７７】

このようにして生成された類似構造に対して、同値類の生成（図８のステップＡ２－６）が行われるが、ここでは、部分構造３ａ－０から生成される一つの類似構造３ａ－２に

行われる変形のみに注目して説明しているため、省略する。

【０１７８】

本実施例における変形では、同義語の置換（図８のステップＡ２－４）が飛ばされるため、図２８（ｃ）の類似構造３ａ－２には、被置換語「高速」が残っている。一方、図２２に示した前記第１、第２の実施例における変形の例では、図２２（ｄ）の類似構造３ａ－１では、被置換語「高速」が代表語「速い」に置換されている。

【０１７９】

本実施例では、このようにして、類似構造生成手段２２が部分構造と類似構造及び同値類の生成を行うことで、図１６の文１の文構造から、図２３に示される同値類が生成され、図１６の文２の文構造から、図２４に示される同値類が生成され、図１６の文３の文構造から図２９に示される同値類が生成される。

【０１８０】

次に、頻出類似パターン検出手段２４は、図２３、図２４、及び図２９に示される同値類の集合から、ステップＣ２で、類似構造判定調整手段２６が指定した属性値の差異を無視しながら頻出パターンの検出を行う（図７のステップＣ４）。

【０１８１】

頻出類似パターン検出手段２４は、要素の少なくとも一つが同一である同値類は同一と判定して、頻出パターンの検出を行う。

【０１８２】

本実施例においては、頻出類似パターン検出手段２４は、類似構造判定調整手段２６からの指定により、どの属性値の差異を無視して類似構造の同一性を判定するかを決定する。

【０１８３】

本実施例では、

「表層格の差異を無視する」、

「付属語情報の差異を無視する」

と動作を制御するように類似構造判定調整手段２６が指定を行ったため、頻出類似パターン検出手段２４は、前記第２の実施例と同様に、類似構造の同一性の判定を行う。

【０１８４】

本実施例においては、図２３、図２４、及び図２９を参照すると、

「類似構造１ａ－１、及び、類似構造２ａ－３」、

「部分構造２ｃ－０、及び、部分構造３ｂ－０」、

「類似構造１ｂ－１、類似構造２ｃ－１、及び、類似構造３ｂ－１」、

「部分構造１ｃ－０、及び、類似構造２ｂ－０」、

「類似構造１ｃ－１、及び、類似構造２ｂ－１」、

「部分構造１ｄ－０、及び、部分構造２ｄ－０」、

「部分構造１ｅ－０、部分構造２ｆ－０、及び、部分構造３ｆ－０」、

「部分構造１ｆ－０、部分構造２ｅ－０、及び、部分構造３ｄ－０」

がそれぞれ頻出類似パターン検出手段２４に同一の構造と判定される。

【０１８５】

頻出類似パターン検出手段２４は、要素の少なくとも一つが同一である同値類は同一と判定するため、

「同値類１ａ、及び、２ａ」、

「同値類１ｂ、２ｃ、及び、３ｂ」、

「同値類１ｃ、及び、２ｂ」、

「同値類１ｄ、及び、２ｄ」、

「同値類１ｅ、２ｆ、及び、３ｆ」、

「同値類１ｆ、２ｅ、及び、３ｄ」

をそれぞれ同一の同値類と判定する。

【０１８６】

本実施例では、前記第１、第２の実施例と同様に、３回以上出現する同値類を頻出パターンとする。

【０１８７】

この場合、

「同値類１ｂ、２ｃ、及び、３ｂ」、

「同値類１ｅ、２ｆ、及び、３ｆ」、

「同値類１ｆ、２ｅ、及び、３ｄ」

が頻出パターンとして検出される。

【０１８８】

最後に、このようにして抽出された頻出パターンを表す構造を、出力装置３に表示する（図７のステップＡ４）。

【０１８９】

本実施例において、出力装置３が出力する頻出パターンの表現は、図３０のようになる。図３０に示すように、本実施例では、前記第１、第２の実施例と同様に、頻出パターンを表す同値類の要素である類似構造を頻出パターンの表現として用いている。

【０１９０】

使用者は、この頻出パターン検出に不満を感じた場合、図７のステップＣ１に戻り、どこまで類似した構造を同一と判定するかの指定の入力を変更することで、再度頻出パターンの検出を行うことができる。

【０１９１】

このようにして、

「同義語の置換による差異については同一と判定しない」

という使用者の指定に基づき、図２３、図２４、図２９において、

「部分構造１ａ－０、部分構造２ａ－０、及び、部分構造３ａ－０」、

「部分構造１ｃ－０、部分構造２ｂ－０、及び、部分構造３ｃ－０」、

「部分構造１ｄ－０、部分構造２ｄ－０、及び、部分構造３ｅ－０」

といった類似した意味を持つが使用者の入力に反する構造を同一と判定せずに、頻出パターン検出行うことで、使用者がどこまで類似した構造を同一と判定するかの調整を行うことができる。

【産業上の利用可能性】

【０１９２】

本発明によれば、コンピュータ上に蓄積される、顧客からの苦情メールやアンケート結果の特徴分析を行う目的に良く用いられるテキストマイニング装置や、テキストマイニング装置をコンピュータに実現するためのプログラムといった用途に適用できる。

【図面の簡単な説明】

【０１９３】

【図１】本発明の第１の実施の形態の構成を示す図である。

【図２】第１の実施の形態の動作を説明するための流れ図である。

【図３】本発明の実施の形態における類似構造生成手段２２の動作を説明するための流れ図である。

【図４】本発明の第２の実施の形態の構成を示す図である。

【図５】本発明の第２の実施の形態の動作を説明するための流れ図である。

【図６】本発明の第３の実施の形態の構成を示す図である。

【図７】本発明の第３の実施の形態の動作を説明するための流れ図である。

【図８】本発明の第３の実施の形態における類似構造生成手段２２の動作を説明するための流れ図である。

【図９】本発明の第４の実施の形態の構成を示す図である。

【図１０】従来の技術の構成を示す図である。

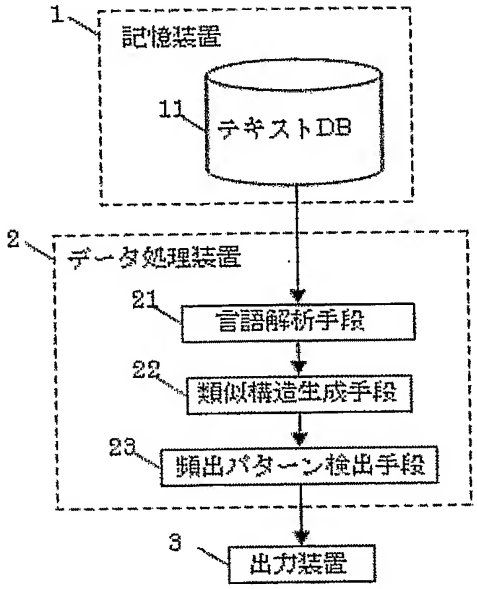
【図１１】属性値を用いる形式で表された「彼は私が本を買ったのを知らない」という文の構文構造の例を示す図である。

【図１２】属性値を用いない形式で表された「彼は私が本を買ったのを知らない」という文の構文構造の例を示す図である
【図１３】連結構造が異なり類似した意味を持つ構造の差異の例を示す図であり、（ａ）は、係り受けの向きの差異、（ｂ）は、係り受けの順序の差異、（ｃ）は、同義語の置換による差異、（ｄ）は、並列の構文構造と意味構造の差異を示す図である。
【図１４】属性値が異なり類似した意味を持つ構造の差異の複数の例を示す図である。（ａ）は、付属語情報の差異を示す図である。（ｂ）は、表層格の差異を示す図である。
【図１５】本発明の第１乃至第３実施例で使用するテキストＤＢ中のテキスト集合の例を示す図である。
【図１６】本発明の第１乃至第３の実施例において、図３に示すテキスト集合中の各テキストを言語解析して得られる文構造の集合を示す図である。
【図１７】本発明の第１乃至第３の実施例において使用する、同義語辞書の構造を示す図である。
【図１８】本発明の第１乃至第３の実施例において、図３のステップＡ２－１における処理を示す図である。
【図１９】本発明の第１乃至第３の実施例において、図３のステップＡ２－２における処理を示す図である。
【図２０】本発明の第１乃至第３の実施例において、図３のステップＡ２－３における処理を示す図である。
【図２１】本発明の第１乃至第３の実施例において、図３のステップＡ２－６における処理を示す図である。
【図２２】本発明の第１、第２の実施例において、類似構造生成手段２２が文３の文構造の全体からなる部分構造３ａ－０の類似構造を生成する処理を示す図である。
【図２３】本発明の第１乃至第３の実施例において文１の文構造から生成される部分構造の同値類を示す図である。
【図２４】本発明の第１乃至第３の実施例において文２の文構造から生成される部分構造の同値類を示す図である。
【図２５】本発明の第１、第２の実施例において、文３の文構造から生成される部分構造の同値類を示す図である。
【図２６】本発明の第１の実施例において、図２３乃至２５に示す同値類の集合から検出される頻出パターンを示す図である。
【図２７】本発明の第２の実施例において、図２３乃至２５に示す同値類の集合から検出される頻出パターンを示す図である。
【図２８】本発明の第３の実施例において、類似構造生成手段２２が文３の文構造の全体からなる部分構造３ａ－０の類似構造を生成する処理を示す図である。
【図２９】本発明の第３の実施例において、文３の文構造から生成される部分構造の同値類を示す図である。
【図３０】本発明の第３の実施例において、図２３、２４及び図２９に示す同値類の集合から検出される頻出パターンを示す図である。

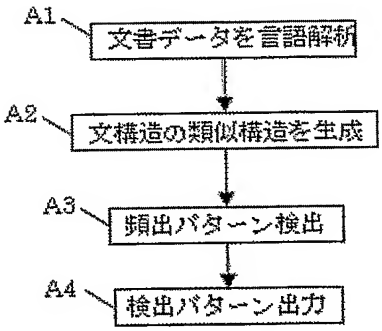
【符号の説明】
【０１９４】
１　記憶装置
２　データ処理装置
３　出力装置
４　データ処理装置
５　データ処理装置
６　入力装置
７　テキストマイニング用プログラム

８　データ処理装置
１１　テキストＤＢ
２１　言語解析手段
２２　類似構造生成手段
２３　頻出パターン検出手段
２４　頻出類似パターン検出手段
２５　類似構造生成調整手段
２６　類似構造判定調整手段

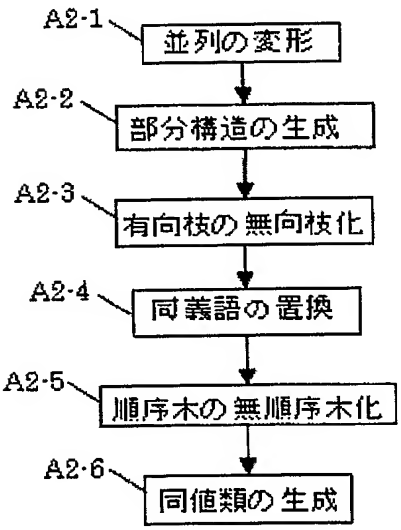
【書類名】図面
【図１】



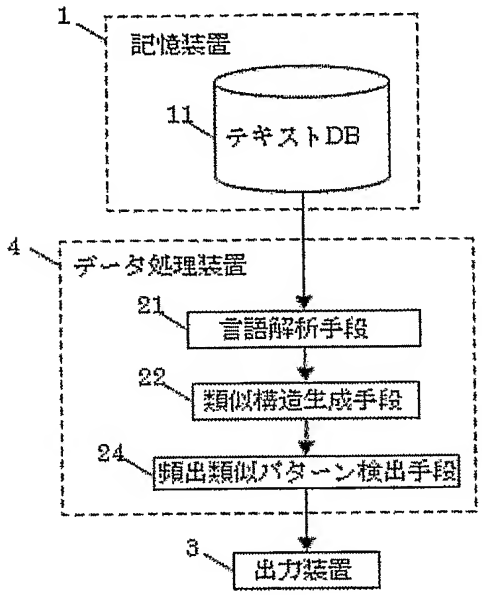
【図２】



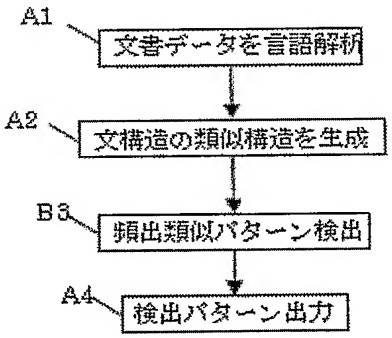
【図３】



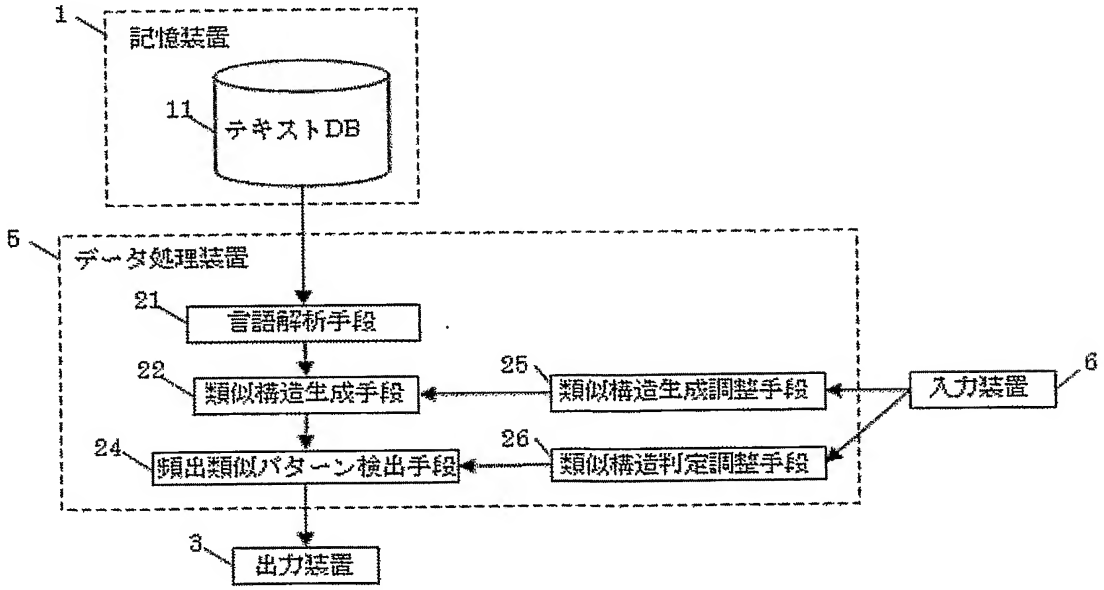
【図４】


1
記憶装置
11
テキストDB
4
データ処理装置
21
言語解析手段
22
類似構造生成手段
24
頻出類似パターン検出手段
3
出力装置

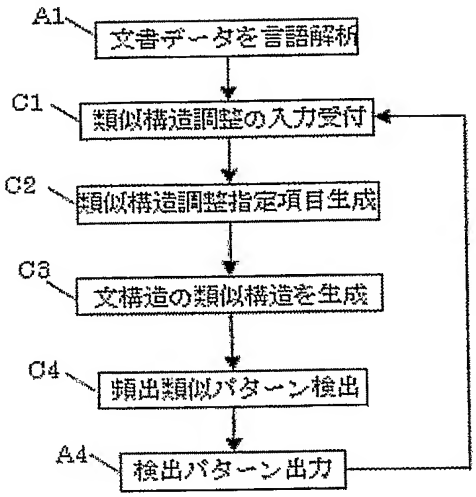
【図５】


A1
文書データを言語解析
A2
文構造の類似構造を生成
B3
頻出類似パターン検出
A4
検出パターン出力

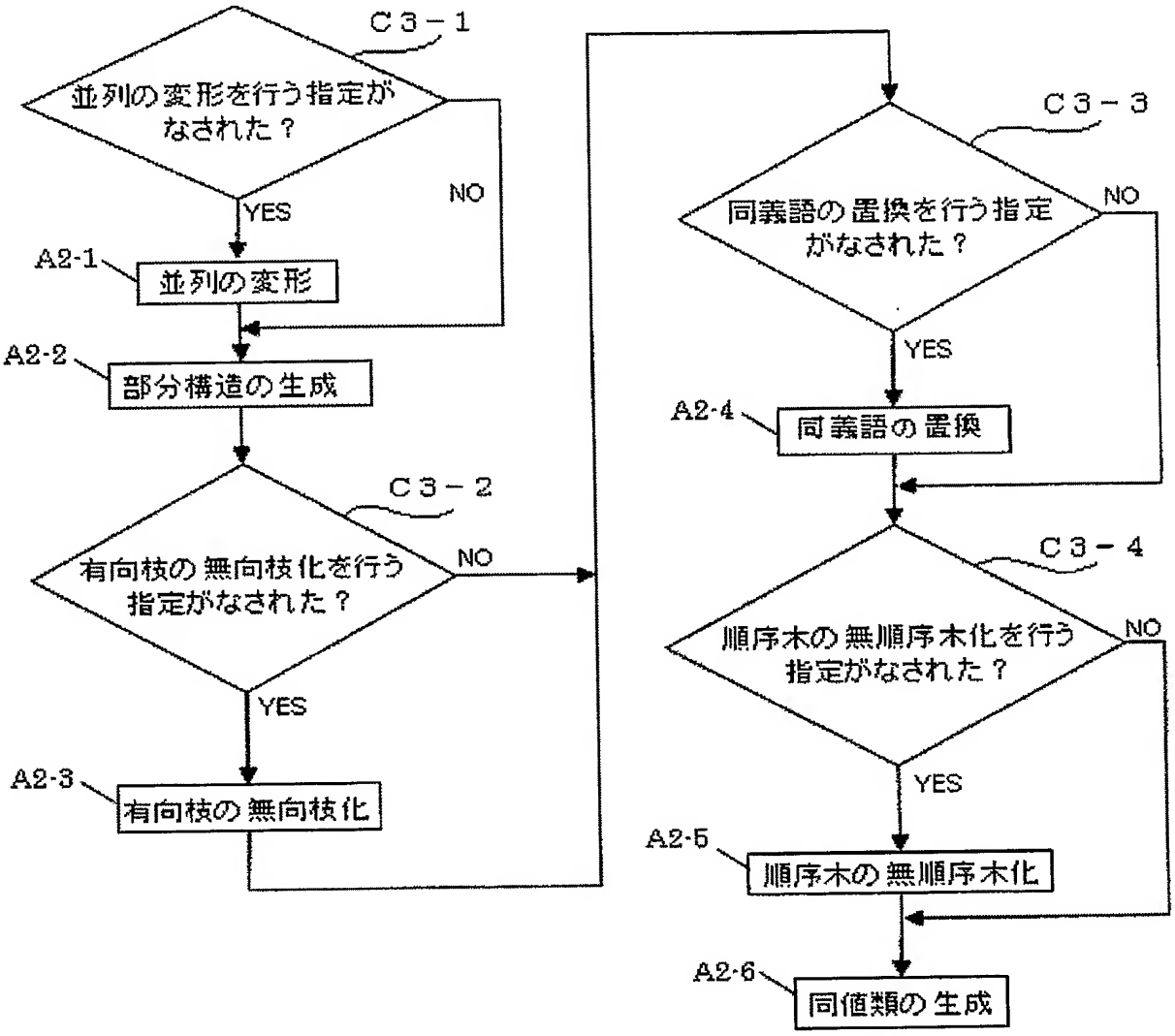
【図６】


1
記憶装置
11
テキストDB
5
データ処理装置
21
言語解析手段
22
類似構造生成手段
25
類似構造生成調整手段
6
入力装置
24
頻出類似パターン検出手段
26
類似構造判定調整手段
3
出力装置

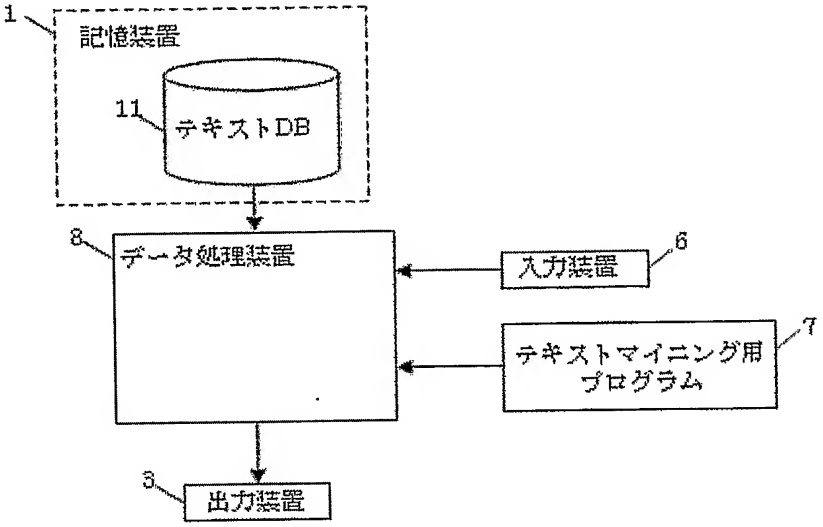
【図７】


A1
文書データを言語解析
C1
類似構造調整の入力受付
C2
類似構造調整指定項目生成
C3
文構造の類似構造を生成
C4
頻出類似パターン検出
A4
検出パターン出力

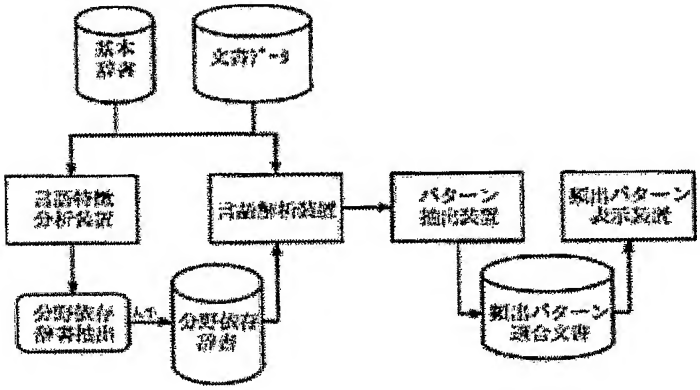
【図８】


C３－１
並列の変形を行う指定がなされた？
YES
NO
A2-1
並列の変形
A2-2
部分構造の生成
C３－２
有向枝の無向枝化を行う指定がなされた？
NO
YES
A2-3
有向枝の無向枝化
C３－３
同義語の置換を行う指定がなされた？
NO
YES
A2-4
同義語の置換
C３－４
順序木の無順序木化を行う指定がなされた？
NO
YES
A2-5
順序木の無順序木化
A2-6
同値類の生成

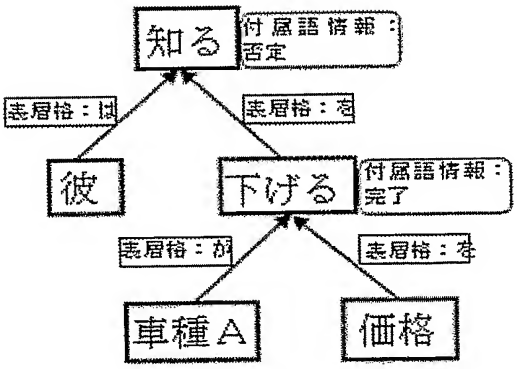
【図９】


1
記憶装置
11
テキストDB
8
データ処理装置
6
入力装置
7
テキストマイニング用
プログラム
3
出力装置

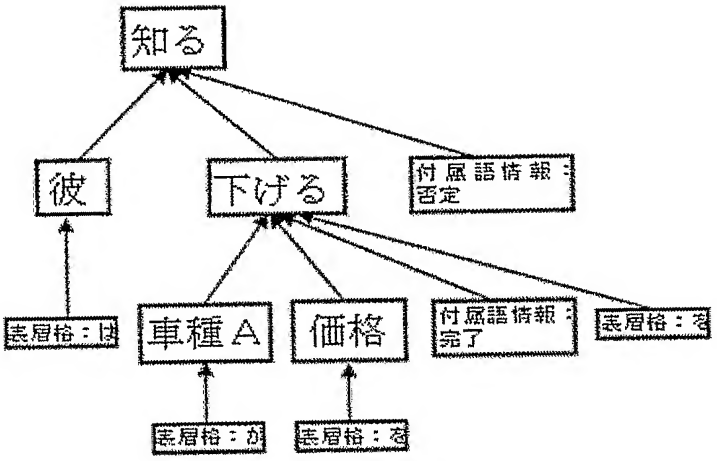
【図１０】


基本
辞書
文書ﾃﾞｰﾀ
言語特徴
分析装置
言語解析装置
パターン
抽出装置
採用パターン
表示装置
分野依存
辞書抽出
分野依存
辞書
抽出パターン
集合文書

【図１１】

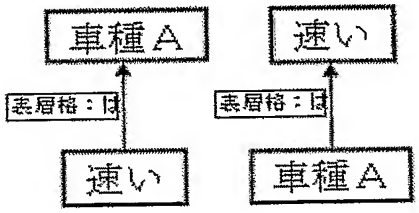

知る
付属語情報：
否定
表層格：は
表層格：を
彼
下げる
付属語情報：
完了
表層格：が
表層格：を
車種Ａ
価格

【図１２】

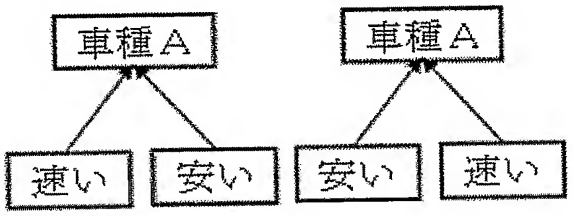


【図１３】

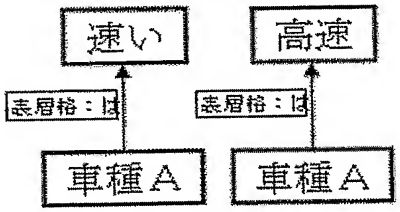
（ａ）「速いのは車種Ａ」と「車種Ａは速い」



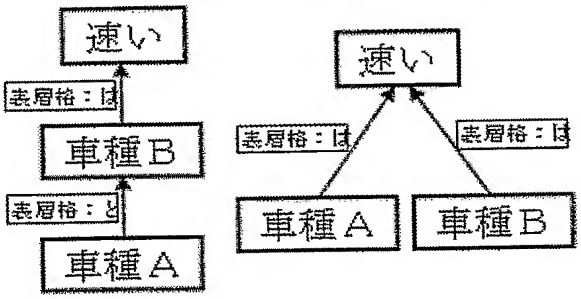
（ｂ）「速く安い車種Ａ」と「安く速い車種Ａ」



（ｃ）「車種Ａは速い」と「車種Ａは高速だ」

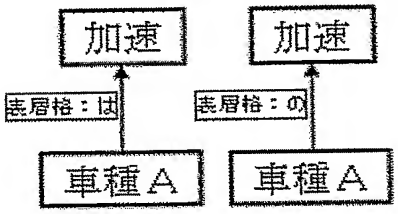


（ｄ）「車種Ａと車種Ｂは速い」の構文構造と意味構造

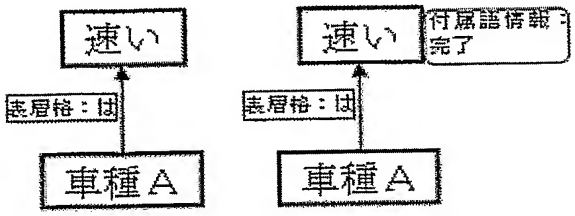


【図１４】

（ａ）「車種Ａは加速」と「車種Ａの加速」



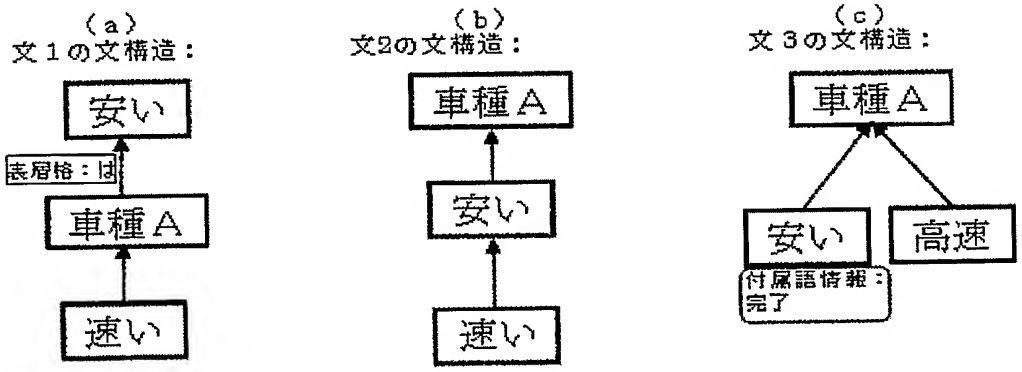
（ｂ）「車種Ａは速い」と「車種Ａは速かった」



【図１５】

文１：速い車種Ａは安い
文２：速く安い車種Ａ
文３：安かった高速な車種Ａ

【図１６】

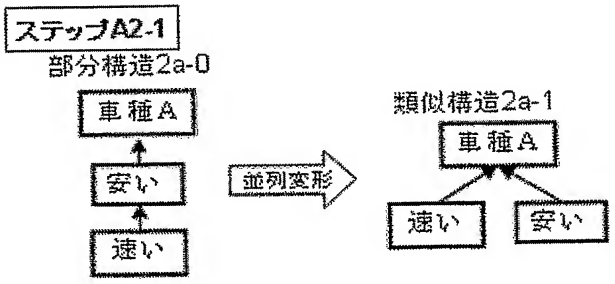


【図１７】

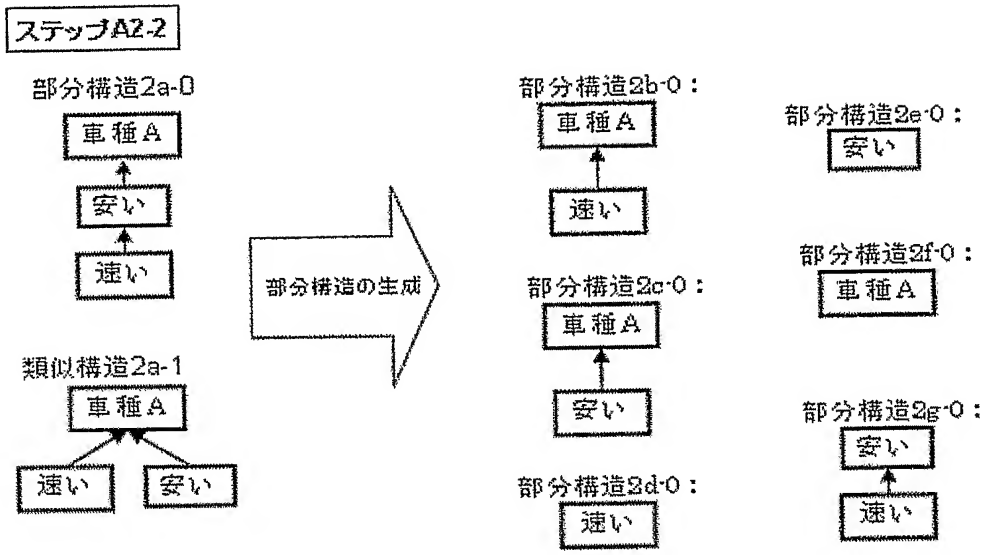
同義語辞書

代表語	被置換後
速い	高速

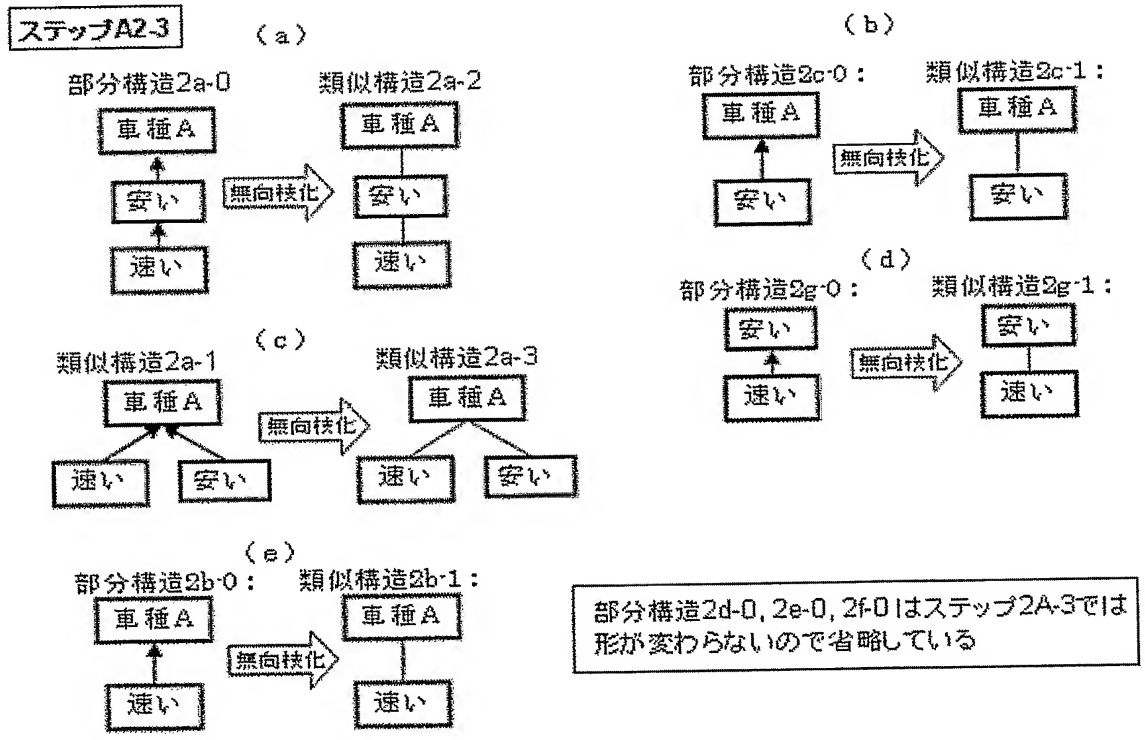
【図１８】



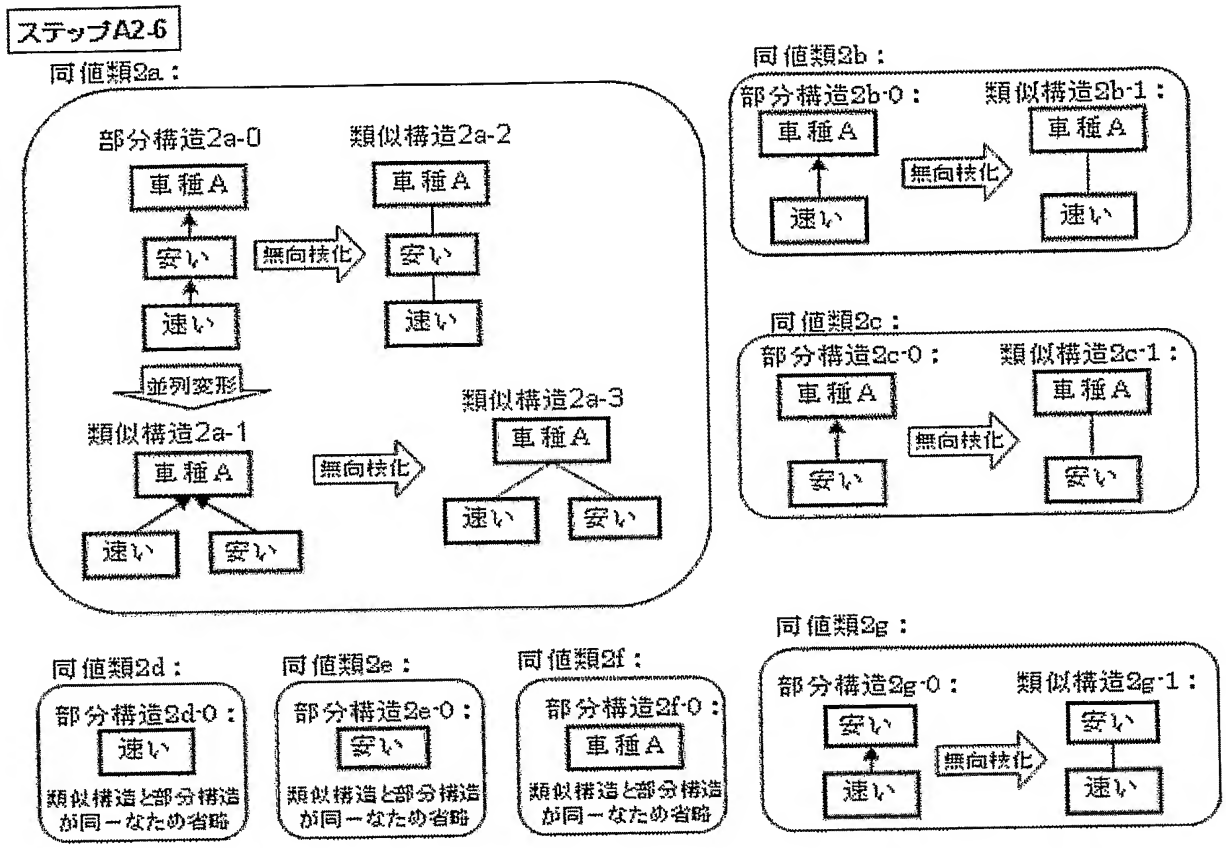
【図１９】


ステップA2-2
部分構造2a-0
車種A
安い
速い
類似構造2a-1
車種A
速い
安い
部分構造の生成
部分構造2b-0：
車種A
速い
部分構造2c-0：
車種A
安い
部分構造2d-0：
速い
部分構造2e-0：
安い
部分構造2f-0：
車種A
部分構造2g-0：
安い
速い

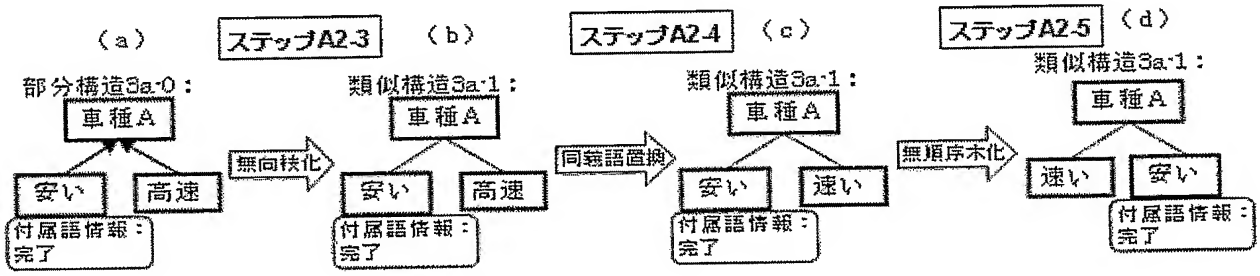
【図２０】


ステップA2-3
（a）
部分構造2a-0
車種A
安い
速い
無向枝化
類似構造2a-2
車種A
安い
速い
（b）
部分構造2c-0：
車種A
安い
無向枝化
類似構造2c-1：
車種A
安い
（c）
類似構造2a-1
車種A
速い
安い
無向枝化
類似構造2a-3
車種A
速い
安い
（d）
部分構造2g-0：
安い
速い
無向枝化
類似構造2g-1：
安い
速い
（e）
部分構造2b-0：
車種A
速い
無向枝化
類似構造2b-1：
車種A
速い
部分構造2d-0, 2e-0, 2f-0はステップ2A-3では形が変わらないので省略している

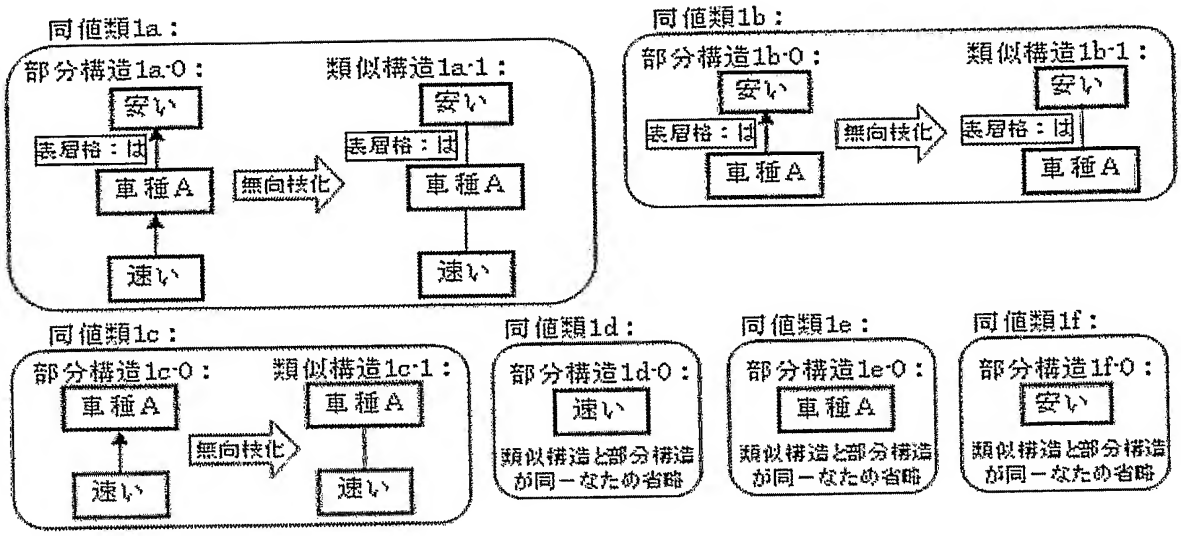
【図２１】


ステップA2-6
同値類2a：
部分構造2a-0
車種A
安い
速い
無向枝化
類似構造2a-2
車種A
安い
速い
並列変形
類似構造2a-1
車種A
速い
安い
無向枝化
類似構造2a-3
車種A
速い
安い
同値類2b：
部分構造2b-0：
車種A
速い
無向枝化
類似構造2b-1：
車種A
速い
同値類2c：
部分構造2c-0：
車種A
安い
無向枝化
類似構造2c-1：
車種A
安い
同値類2d：
部分構造2d-0：
速い
類似構造と部分構造が同一なため省略
同値類2e：
部分構造2e-0：
安い
類似構造と部分構造が同一なため省略
同値類2f：
部分構造2f-0：
車種A
類似構造と部分構造が同一なため省略
同値類2g：
部分構造2g-0：
安い
速い
無向枝化
類似構造2g-1：
安い
速い

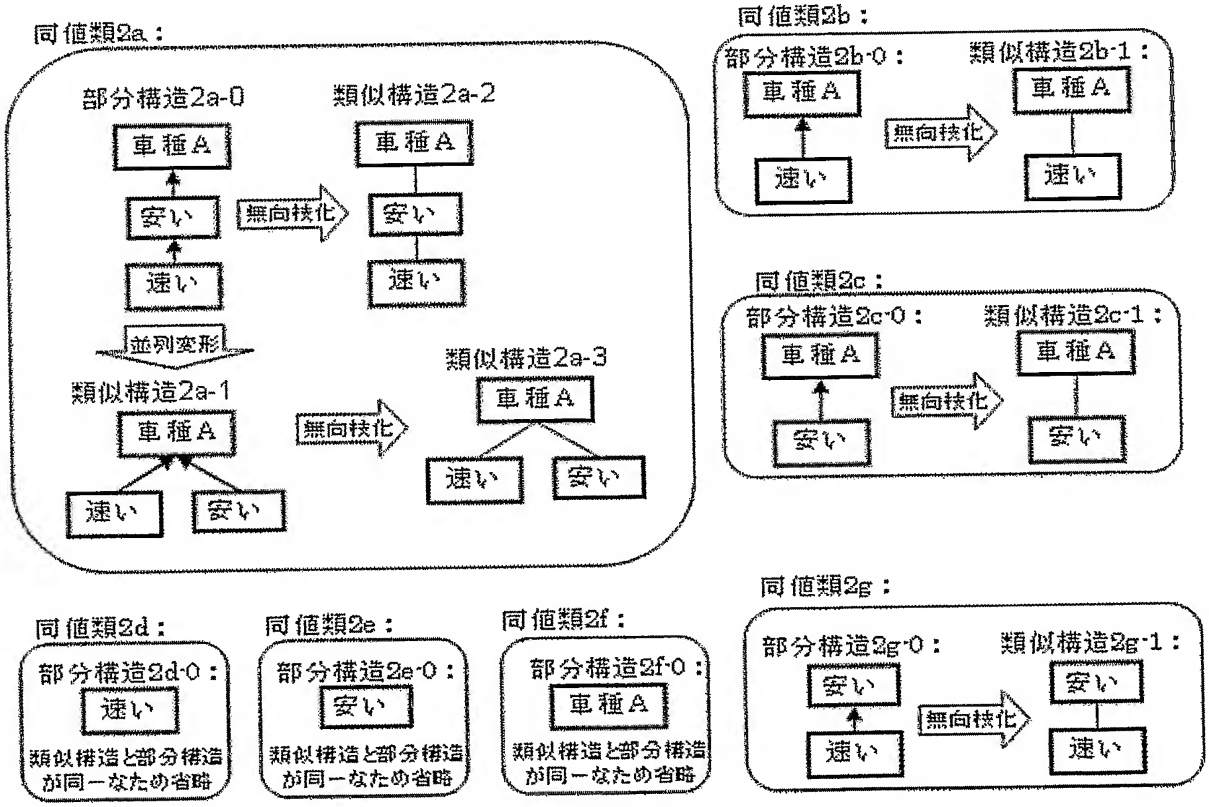
【図２２】


（a）
部分構造3a-0：
車種A
安い
高速
付属語情報：完了
ステップA2-3
無向枝化
（b）
類似構造3a-1：
車種A
安い
高速
付属語情報：完了
ステップA2-4
同義語置換
（c）
類似構造3a-1：
車種A
安い
速い
付属語情報：完了
ステップA2-5
無順序木化
（d）
類似構造3a-1：
車種A
速い
安い
付属語情報：完了

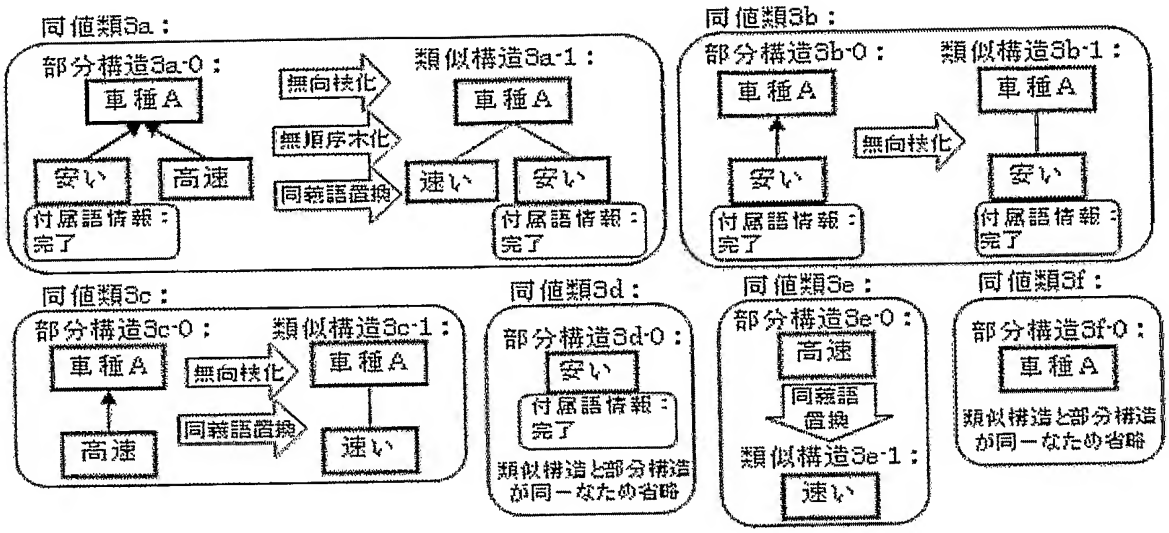
【図２３】


同値類1a：
部分構造1a-0：
安い
表層格：ｶﾅ
車種A
速い
無向枝化
類似構造1a-1：
安い
表層格：ｶﾅ
車種A
速い
同値類1b：
部分構造1b-0：
安い
表層格：ｶﾅ
車種A
無向枝化
類似構造1b-1：
安い
表層格：ｶﾅ
車種A
同値類1c：
部分構造1c-0：
車種A
速い
無向枝化
類似構造1c-1：
車種A
速い
同値類1d：
部分構造1d-0：
速い
類似構造と部分構造が同一なため省略
同値類1e：
部分構造1e-0：
車種A
類似構造と部分構造が同一なため省略
同値類1f：
部分構造1f-0：
安い
類似構造と部分構造が同一なため省略

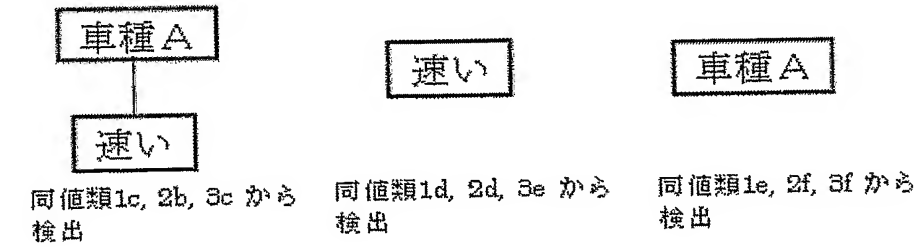
【図２４】


同値類2a：
部分構造2a-0
車種A
安い
速い
無向枝化
類似構造2a-2
車種A
安い
速い
並列変形
類似構造2a-1
車種A
速い
安い
無向枝化
類似構造2a-3
車種A
速い
安い
同値類2b：
部分構造2b-0：
車種A
速い
無向枝化
類似構造2b-1：
車種A
速い
同値類2c：
部分構造2c-0：
車種A
安い
無向枝化
類似構造2c-1：
車種A
安い
同値類2d：
部分構造2d-0：
速い
類似構造と部分構造が同一なため省略
同値類2e：
部分構造2e-0：
安い
類似構造と部分構造が同一なため省略
同値類2f：
部分構造2f-0：
車種A
類似構造と部分構造が同一なため省略
同値類2g：
部分構造2g-0：
安い
速い
無向枝化
類似構造2g-1：
安い
速い

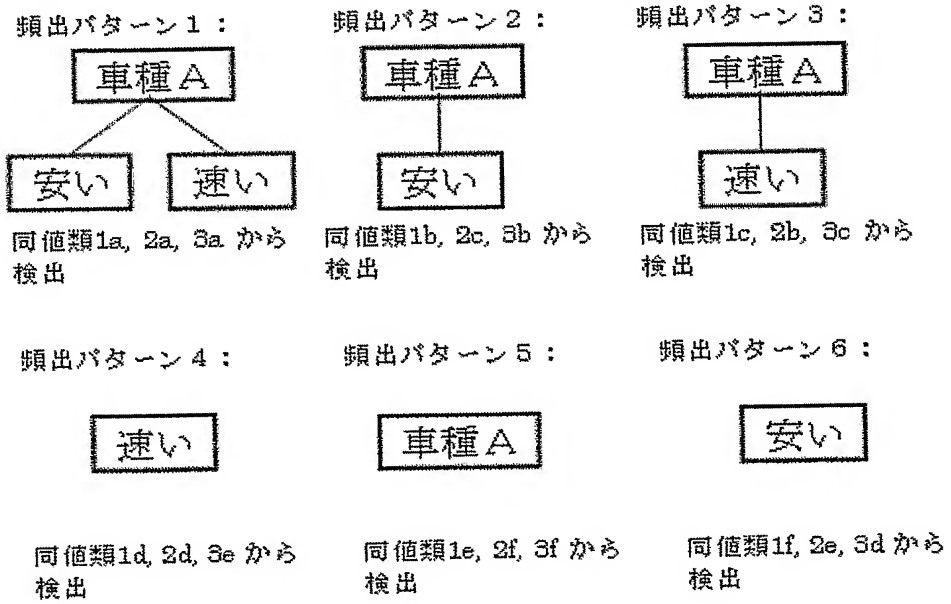
【図２５】


同値類3a：
部分構造3a-0：
車種A
安い
高速
付属語情報：完了
無向枝化
無順序木化
同義語置換
類似構造3a-1：
車種A
速い
安い
付属語情報：完了
同値類3b：
部分構造3b-0：
車種A
安い
付属語情報：完了
無向枝化
類似構造3b-1：
車種A
安い
付属語情報：完了
同値類3c：
部分構造3c-0：
車種A
高速
無向枝化
同義語置換
類似構造3c-1：
車種A
速い
同値類3d：
部分構造3d-0：
安い
付属語情報：完了
類似構造と部分構造が同一なため省略
同値類3e：
部分構造3e-0：
高速
同義語置換
類似構造3e-1：
速い
同値類3f：
部分構造3f-0：
車種A
類似構造と部分構造が同一なため省略

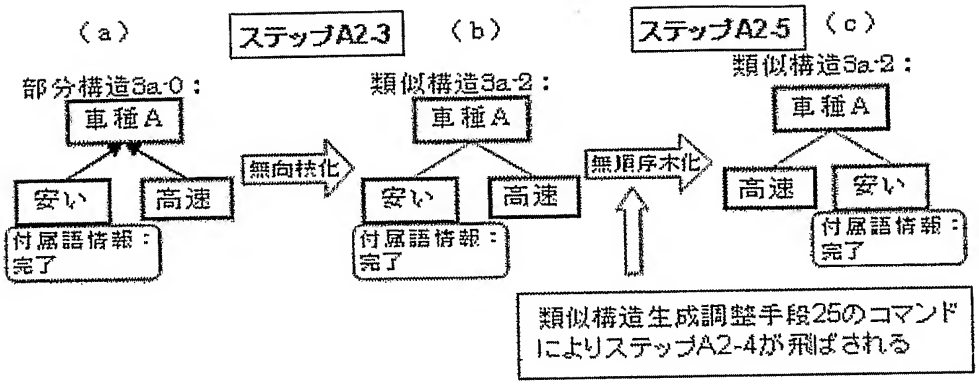
【図２６】


頻出パターン１：
車種A
速い
同値類1c, 2b, 3c から検出
頻出パターン２：
速い
同値類1d, 2d, 3e から検出
頻出パターン３：
車種A
同値類1e, 2f, 3f から検出

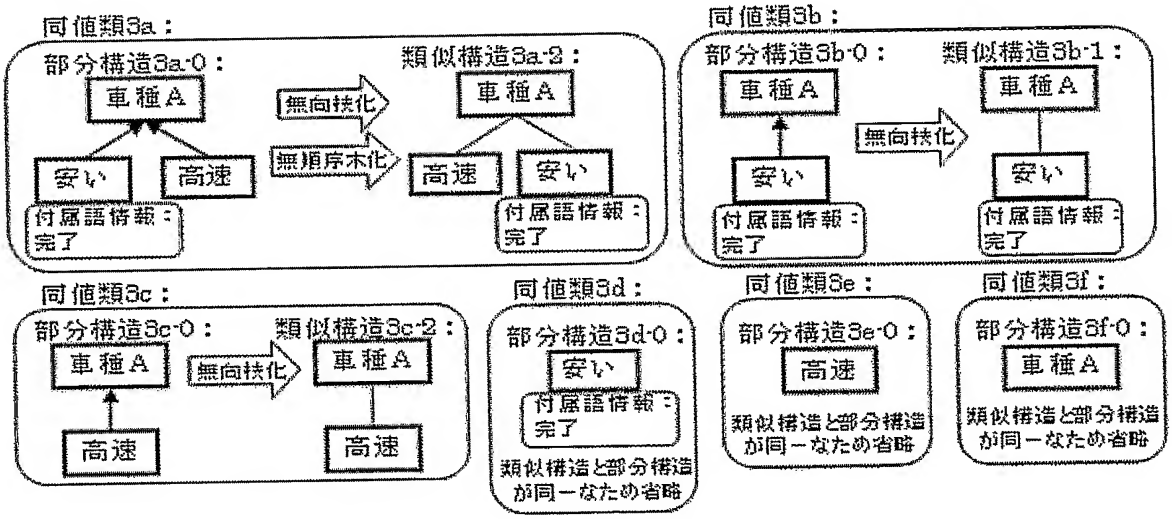
【図２７】


頻出パターン１：
車種A
安い
速い
同値類1a, 2a, 3a から検出
頻出パターン２：
車種A
安い
同値類1b, 2c, 3b から検出
頻出パターン３：
車種A
速い
同値類1c, 2b, 3c から検出
頻出パターン４：
速い
同値類1d, 2d, 3e から検出
頻出パターン５：
車種A
同値類1e, 2f, 3f から検出
頻出パターン６：
安い
同値類1f, 2e, 3d から検出

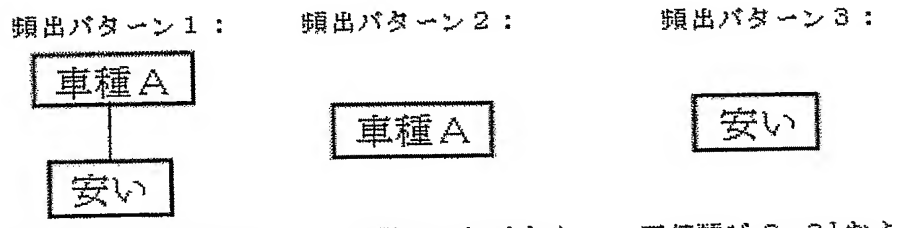
【図２８】


（ａ）
ステップA2-3
（ｂ）
ステップA2-5
（ｃ）
部分構造3a-0：
車種Ａ
安い
高速
付属語情報：完了
無向枝化
類似構造3a-2：
車種Ａ
安い
高速
付属語情報：完了
無順序木化
類似構造3a-2：
車種Ａ
高速
安い
付属語情報：完了
類似構造生成調整手段25のコマンドによりステップA2-4が飛ばされる

【図２９】


同値類3a：
部分構造3a-0：
車種Ａ
安い
高速
付属語情報：完了
無向枝化
無順序木化
類似構造3a-2：
車種Ａ
高速
安い
付属語情報：完了
同値類3b：
部分構造3b-0：
車種Ａ
安い
付属語情報：完了
無向枝化
類似構造3b-1：
車種Ａ
安い
付属語情報：完了
同値類3c：
部分構造3c-0：
車種Ａ
高速
無向枝化
類似構造3c-2：
車種Ａ
高速
同値類3d：
部分構造3d-0：
安い
付属語情報：完了
類似構造と部分構造が同一なため省略
同値類3e：
部分構造3e-0：
高速
類似構造と部分構造が同一なため省略
同値類3f：
部分構造3f-0：
車種Ａ
類似構造と部分構造が同一なため省略

【図３０】


頻出パターン１：
車種Ａ
安い
同値類1b, 2c, 3b から検出
頻出パターン２：
車種Ａ
同値類1e, 2f, 3f から検出
頻出パターン３：
安い
同値類1f, 2e, 3d から検出

【書類名】要約書
【要約】
【課題】
意味的に類似した文構造を同一と判定しながら頻出パターンの検出を行うテキストマイニングを可能にする。さらに、どこまで類似した構造を同一と判定して頻出パターン検出を行うかを使用者が調整可能なテキストマイニングを可能にする装置、方法、プログラムの提供。
【解決手段】
言語解析手段２１はテキストＤＢ１１から読み込んだ各テキストの解析を行い解析結果として文構造を生成し、入力装置６は文構造の差異の種別毎に同一構造と判定するか否を決定するための入力と属性値の種別毎に値の差異を無視するか否かを決定するための入力を受け、類似構造生成調整手段２５は入力装置からの入力より文構造の差異の種別毎に同一構造と判定するか否かを指定する指定項目を生成し、類似構造判定調整手段２６は入力装置６からの入力より属性値の種別毎に値の差異を無視するか否かを指定する指定項目を生成し、類似構造生成手段２２は類似構造生成調整手段２５からの指定項目に従い言語解析手段２１が得た文構造の集合中の各構造の部分構造の類似構造を生成し生成した各類似構造を夫々の生成元の部分構造の同値類とし、頻出類似パターン検出手段２４は類似構造判定調整手段２６より与えられた指定項目に従い属性値を無視し類似構造生成手段２２からの同値類の集合より頻出パターンを検出し出力装置３に出力する。
【選択図】
図６

特願２００４－０７９０７７

出　願　人　履　歴　情　報

識別番号　［０００００４２３７］

１．変更年月日　１９９０年　８月２９日
［変更理由］　新規登録
住　所　東京都港区芝五丁目７番１号
氏　名　日本電気株式会社